# Relationale Messung berufsrelevanter Kompetenzen und deren Visualisierung mit Hilfe der Multidimensionalen Skalierung

Abhandlung
zur Erlangung der Doktorwürde
der Philosophischen Fakultät
der Universität Zürich

vorgelegt von
**Tony Zuber**
von Wattwil (SG)

Angenommen im Herbstsemester 2010 auf Antrag von
Herrn Prof. Dr. Damian Läge und
Herrn Prof. Dr. Martin Kleinmann

Zentralstelle der Studentenschaft der Universität Zürich, 2011

# Vorwort

Als ich mich 2006 entschloss, eine berufsbegleitende Dissertation in Angriff zu nehmen, ahnte ich nicht, welche Folgen dies für mich sowohl beruflich als auch privat haben würde. Damals standen das Interesse am Thema und die wahrgenommene Relevanz der vorliegenden Untersuchung für die praktische Eignungsdiagnostik im Vordergrund. Die Herausforderung, die beruflichen und wissenschaftlichen Anforderungen unter einen Hut zu bringen, hat mich jedoch schnell eingeholt und mich gezwungen, Prioritäten zu setzen. Diese gingen oft zu Gunsten meines Arbeitgebers und zu Lasten meines persönlichen Ziels, die wissenschaftliche Arbeit möglichst in einem Zug, ohne grössere Unterbrechungen und Verzögerungen, durchzuziehen. Ich habe wiederholt daran gezweifelt, ob ich die wissenschaftliche Arbeit unter dem wachsenden Druck der steigenden beruflichen Anforderungen tatsächlich beenden könnte. Insofern gebührt der erste und ganz besondere Dank zwei meiner Vorgesetzten, Herr Martin Lüthy und Herrn Rolf Curschellas, die mich stets in meinem Vorhaben unterstützten und mir die nötige Handlungsfreiheit schufen und Flexibilität entgegenbrachten, um die Arbeit zu einem Abschluss zu bringen. Auch meinen drei Mitarbeiterinnen, Frau Patrizia May, Frau Claudia Burkard und Frau Anne Forster gebührt mein Dank. Sie ermöglichten durch ihre hohe Selbständigkeit und ihr Verständnis für meine persönlichen Ziele die parallele Arbeit an der Universität. Alle diese Menschen haben an meine Fähigkeiten geglaubt und mir den Rücken gestärkt. Weiter gilt mein aufrichtiger Dank den Führungskräften und Personalverantwortlichen des Axpo Konzerns, die mich bei der Durchführung der Untersuchung und der Datenerhebung halfen. Sie erlaubten mir, im Rahmen meiner Befragungen, die empirische Grundlage für die Arbeit zu schaffen.

Ganz besonders herzlich möchte ich mich bei Herrn Prof. Dr. Damian Läge bedanken. Herr Läge war als Betreuer und Erstgutachter stets für mich erreichbar, unterstützend und fördernd. Durch sein genuines Interesse am Thema und seinem unermüdlichen Willen, die sozialwissenschaftlichen Methoden in didaktisch hochstehender Form zu vermitteln, hat er mich nicht nur fachlich unterstützt, sondern auch meine methodische Kompetenz erweitert. Seine Diskussionsbereitschaft bis in die Abendstunden und seine offene Haltung gegenüber neuen Ideen, resultierten in neuen

wissenschaftlichen Erkenntnissen und machten den Abschluss der Arbeit erst möglich.
Herzlicher Dank gilt ferner Herrn Prof. Dr. Martin Kleinmann für seine Unterstützung als Zweitgutachter sowie Herrn Dr. Stefan Ryf für die Programmierung der online gestützten Erhebungsinstrumente.

Nicht zuletzt möchte ich mich bei meiner Frau Asa Massleberg bedanken, die ihre eigenen Bedürfnisse oft zurücksteckte und stets die richtigen Worte fand, um mich in meinem Vorhaben zu unterstützen.
Bei meiner alleinerziehenden Mutter Marlis Zuber bedanke ich mich zum Schluss für ihr fürsorgliches Engagement während meiner gesamten Ausbildungszeit, von der Primarschule bis zum Doktorat. Es liegt auf der Hand, dass ohne sie diese Doktorarbeit nicht möglich gewesen wäre.

# I. Inhaltsverzeichnis

# 1 Zusammenfassung / Abstract

Die vorliegende Arbeit beschäftigt sich mit der Messung berufsbezogener Kompetenzen, sofern diese in der Persönlichkeit von Personen begründet liegen. Modelliert werden dabei leistungs- und kompetenzrelevante Daten aus dem Feld der Managementberufe. Der erste Schwerpunkt liegt im Bereich der Messung, indem (intervallskalierte) Rangordnungen von Kompetenzitems durch graphische Drag&Drop-Umgebungen erstellt und gängigen Fragebögen (auf Mehrpunktskalen) gegenübergestellt werden. Solche Forced-Choice-Verfahren, welche man zur Klasse der ipsativen Messinstrumente zählen kann, erfordern ein anderes Auswertemodell als normative Messinstrumente, auf die die klassische Testtheorie anwendbar ist. Deswegen wird – und das ist der zweite methodische Schwerpunkt der Arbeit – ein MDS-basiertes relationales Vorgehen für die Modellierung der Daten vorgeschlagen und anhand von drei Datensätzen aus einem Grossunternehen der Energiewirtschaft überprüft. Es zeigt sich, dass den Vorteilen der Rangordnungen (deutlich höhere Effizienz, Transparenz in der Messung und Unterdrückung von Beurteilertendenzen) kein Nachteil in der Qualität der Modellierung der Daten gegenüber steht, sondern dass das effizientere Forced-Choice-Verfahren ähnliche Gütemasse erreicht wie die in der Praxis wesentlich weiter verbreiteten normativen Verfahren. Bei der MDS werden zudem die relationalen Beziehungen zwischen einzelnen Kompetenzprofilen berücksichtigt und die gesamte Struktur der Daten anhand zweidimensionaler euklidischer Karten abgebildet. Dadurch können Personalbeurteilungsdaten in einer neuen und sowohl für die Wissenschaft als auch für die Praxis zweckmässigen Weise ganzheitlich dargestellt und interpretiert werden.

The current work explores the measurement of managerial competencies. The first focus is on the type of measurement. A forced-choice procedure, where the rater puts competency items through drag&drop into an (interval scaled) rankorder is compared to common (multipoint) questionnaires. The data collected through a forced-choice procedure, which belong to the class of ipsative measurement, requires different methods of analysis than normative instruments, where classical test theory can be applied. Therefore – and this is the second methodological focus of this work – a multidimensional scaling technique is suggested in order to modellize the relational structure of the data collected in three different samples of a large utility company. It is shown, that the advantages of a rankorder-methodoligy (highly efficient, transparency of what is measured and avoidance of rater response tendencies) do not compete with the disadvantage of a reduced quality in modelling the data. The introduced forced-choice application does show similar reliability measures as the much more frequently used normative instruments. In addition, when applying MDS, the relational structure of all objects (i.e. competency profiles) is considered and visualized through two-dimensional Euclidian maps. This allows research and practice a new holistic form of visualising and interpreting human performance data.

# 2 Einleitung

Wir leben in einer Zeit, in der Wissen zur zentralen Voraussetzung gesellschaftlicher Entwicklung und zur wichtigsten Produktivkraft geworden ist. Wissen als strategischer Wettbewerbsfaktor spielt im Wertschöpfungsprozess eines Unternehmens eine bedeutungsvolle Rolle (Wunderer & Dick, 2007). Folglich hängt der Erfolg eines jeden Unternehmens auch wesentlich von seinen Mitarbeitern ab. In diesem Zusammenhang stellt der Mitarbeiterstamm eine Form von Kapital, das „Humankapital“, dar (Fersch, 2002). Dabei gilt es, wie bei jedem anderen Kapital auch, dessen Wert nicht nur aufrecht zu erhalten, sondern auch in qualitativer Weise zu steigern. Dafür ist ein zielorientiertes Personalmanagement der Mitarbeiterpotentiale, welches einen kontinuierlichen Wissenstransfer nachhaltig unterstuützt, ein wesentlicher Erfolgsfaktor für das Unternehmen (Berthel, 2000). In wie weit ein Unternehmen seine Wettbewerbsfähigkeit aufrecht erhalten kann, hängt somit zunehmend auch davon ab, wie gut es einem Unternehmen gelingt, die Auswahl, Entwicklung und Motivation seiner Mitarbeiter zu optimieren. Die korrekte Einschätzung, Entwicklung und Förderung von Beschäftigten gewinnt dadurch zunehmend an Bedeutung (Fersch, 2002). Daraus resultiert auch ein wachsendes Interesse an der Personalbeurteilung als Instrument des modernen Personalmanagements.

Im Themenbereich der Personalbeurteilung werden in Praxis und Literatur die verschiedensten Begriffe verwendet, welchen teils unterschiedliche, teils synonyme Bedeutungen zugeordnet werden. Zu diesen Begriffen zählen u. a. die

- Leistungsbeurteilung
- Mitarbeiterbeurteilung
- Personalbeurteilung
- Potentialbeurteilung
- Kompetenzbeurteilung

Die folgende Abbildung soll einen raschen Überblick verschaffen und zeigt, dass sowohl Inhalt und Zweck der verschiedenen Beurteilungen unterschiedlich sein können.

| **Art der Beurteilung** | **Inhalt der Beurteilung** | **Zweck** |
|---|---|---|
| Leistungsbeurteilung (Vergangenheitsorientiert) | - Zielerreichung der vorgegeben Leistungsziele<br>- Arbeitsverhalten | - Salärbestimmung<br>- Beförderungen<br>- Leistungssteigerung |
| Potentialbeurteilung (Zukunftsorientiert) | Beurteilung von Persönlichkeitseigenschaften wie Gewissenhaftigkeit, Intelligenz, Extraversion etc. | - Personalauswahl<br>- Potentialeinschätzung zur Abklärung von Führungsqualitäten |
| Kompetenzbeurteilung (Vergangenheits- und / oder Zukunftsorientiert) | Beurteilung von verhaltensnahen und berufsrelevanten fachspezifischen und fachübergreifenden Kompetenzen | - Personalentwicklung<br>- Personalauswahl |

Die Kompetenzbeurteilung, die in dieser Arbeit im Fokus steht, ist aus der neueren Competency Bewegung entstanden, welche mit dem Paradigmenwechsel von McClelland (1973) durch seinen Artikel Testing for Competence Rather Than for „Intelligence" bereits früh initiiert wurde. Laut Sarges (2006) haben allein in der letzten Dekade tausende von Firmen weltweit Competency-Untersuchungen in Auftrag gegeben, die als Basis für Entscheidungen über Einstellungen, Trainings, Beförderungen und andere HR-Aktivitäten dienen. Competencies seien laut Sarges inzwischen im Beratergeschäft derart „in“, dass sich letzlich kaum mehr ein HR-Berater dem entziehen dürfte – zumal sich mit Competency-Models nach wie vor guter Umsatz generieren lässt. Dies alles gilt natürlich nicht nur für die USA, sondern – wegen der Internationalität der weltweit operierenden Konzerne – auch für Europa, und das wohl auch noch für eine geraume Weile: „Competence – and its role in achieving peak performance – remains one of the hot issues in business today“ (Zwell, 2000). Auch Giber, Carter & Goldsmith (2009) heben die zentrale Bedeutung von Kompetenzen für die Umsetzung der Strategie von Organisationen hervor: „The future leadership requirement analysis determines the critical competencies required of leaders to deliver on the organization´s future strategy. Once these leadership competencies are identified, they serve as the foundation for the relevant HR

processes that must be aligned with the leadership development tools and processes in order to deliver the leaders required to execute future strategy“.

Die vorliegende Arbeit nimmt diesen Trend auf und baut die empirischen Untersuchungen auf einem für ein Energieunternehmen entwickelten Kompetenzmodell auf. Die Messinstrumente, die für die vielen Competency Models der letzten 20 Jahre angeboten wurden, sind allermeist nicht nach den nötigen psychometrischen Standards konstruiert und evaluiert worden (Barrett & Depinet, 1991; Kurz & Bartram, 2002). Insgesamt erfahren Kompetenzmodelle vor allem in der Praxis hohe Beliebtheit und sind in der Forschung der Angewandten Arbeits- und Organisationspsychologie im Vergleich dazu noch nicht auf allzu grosses Interesse gestossen. Die Forschung der letzten 30 Jahre hat weniger die Kompetenzmessung, als vielmehr die klassische Leistungsbeurteilung und Persönlichkeitsdiagnostik im Fokus gehabt. Einen Überblick über die zahlreichen empirischen Studien zum Thema Personalbeurteilung zu gewinnen, ist in der globalen Wissensgesellschaft ohnehin keine einfache Aufgabe. Sie reicht von psychometrischen Untersuchungen zu Gütekriterien von diversen Leistungs- und Persönlichkeitsfragebögen, über die prognostische Validität von Leistungsbeurteilungen und Assessment Centern bis zu empirischen Studien von der Beurteilerübereinstimmung diverser Beurteilergruppen (Selbstbeurteilungen, Mitarbeiterbeurteilung, Vorgesetztenbeurteilung, Gleichgestelltenbeurteilung). Es ist auch nicht das erklärte Ziel dieser Arbeit, einen umfassenden Überblick über den aktuellen Stand der Forschung zu geben. Dazu haben sich schon viele fleissige Wissenschaftler die Mühe gemacht, das vorhandene Wissen systematisch zusammen zu tragen und strukturiert darzustellen. Der interessierte Leser sei an dieser Stelle an Standardwerke[1] verwiesen.

Das Hauptinteresse der vorliegenden Arbeit liegt in der neuartigen Messung und Visualisierung von leistungsrelevantem Verhalten mittels einer in der gängigen

[1] Ferris, G.H (2006). Research in Personnel and Human Resources Management, Volume 19;. Greenwich, CT: JAI Press.

Viswesvaran, C. (2002) Assessment of individual job performance: A review of the past century and a look ahead. In Anderson, N. & Ones, D. S. (Eds.) Handbook of industrial, work and organizational psychology, Volume 1: Personnel Psychology. Sage Publications: London. 110-126.

Motowidlo, S. (2002). Job performance. In W. C. Borman, D. R. Ilgen, R. J. Klimoski, & I. B. Weiner (Eds.), Handbook of Psychology, Industrial and Organizational Psychology (Vol. 12, pp. 39-53). New York: John Wiley and Sons.

Forschungstradition der Personaldiagnostik noch nie angewandten Methodik[2], der der Multidimensionalen Skalierung (MDS). In der hier vorliegenden Arbeit sprechen wir jedoch von der Nonmetrischen Multidimensionalen Skalierung (NMDS), ein iteratives Verfahren der MDS, welches in der Literatur vor allem bei ordinalem Skalenniveau oft auch als NMDS bezeichnet wird (Borg, Groenen, Mair, 2010).

Der Forschungsschwerpunkt liegt auf dem Vergleich von gängigen Fragebogenformaten mehrstufiger Likertskalen und Forced-Choice-Verfahren. Die zentrale Hypothese dabei lautet, dass Beurteiler bei den gängigen kognitiv überfordert sind, da sie einerseits die Profilbildung über mehrere Kompetenz-Dimensionen und andererseits die absolute Profilhöhe im Vergleich zu anderen Personen in die Bewertung integrieren müssen. Der Autor der vorliegenden Arbeit postuliert demnach ein Forced-Choice-Verfahren zur kognitiven Entlastung, dass einerseits den Prozess der Leistungsbeurteilung in zwei Schritte teilt, nämlich in den Prozess der Profilbildung und der Messung der absoluten Profilhöhe, und andererseits eine praktikable und ökonomische Alternative zu den gängigen langatmigen Leistungsbeurteilungsbögen sowie Persönlichkeits- und Kompetenzinventare bietet.

Unter Anwendung eines im Rahmen dieser Arbeit entwickelten Kompetenzmodells für ein grösseres Schweizerisches Energieunternehmen wird im Hinblick auf die oben formulierte zentrale Hypothese folgenden Forschungsfragen nachgegangen, welche in drei Bereiche gegliedert sind und in sieben wissenschaftlichen Beiträgen erörtert werden. Dabei wird in jedem Beitrag der aktuelle Forschungsstand zur jeweiligen Fragestellung aufgeführt.

[2] Dem Autor sind bis dato keine Studien bekannt, bei denen die Nonmetrische Multidimensionale Skalierung als Methode zur Messung der Gütekriterien von Personalbeurteilungsdaten angewendet wurde.

| **Forschungsfrage** | **Forschungsbeiträge** | **Datensatz** |
|---|---|---|
| 1. Beurteilung der Qualität verschiedener Skalierungsverfahren (Forced-choice vs. Likert type items) | **Forschungsbeitrag I:** The quality of competency based performance ratings in a practical context | Datensatz A (N=15) |
| | **Forschungsbeitrag II:** Rater- and rateeeffect depending on the rating format | Datensatz A (N=15) |
| | **Forschungsbeitrag III:** Measuring overall performance through absolute and relative rating format | Datensatz B (N=45) |
| | **Forschungsbeitrag IV:** Relationale Kompetenz-messung und deren Visualisierung | Datensatz C (N=28) |
| | **Forschungsbeitrag V:** Relationale Kompetenzmessung und dessen Anwendbarkeit auf unterschiedliche Kompetenzmodelle | Datensatz C (N=28) |
| 2. Die Übereinstimmung zwischen Selbst- und Fremdurteil bei ipsativer Messung | **Forschungsbeitrag VI** Kongruenz von Selbst- und Fremdbild bei ipsativer Kompetenzmessung | Datensatz C (N=28) |
| 3. Die Überprüfung der Reliabilität eines bevorzugten Skalierungsverfahrens | **Forschungsbeitrag VII:** Forced-choice measurement: Reliability of competency scaling depending on the rating and item format | Datensatz C (N=28) |

Die sieben Forschungsbeiträge beruhen auf drei verschiedenen Datensätzen, welche im Rahmen dieser berufsbegleitenden Dissertation mit unterschiedlichen Stichproben innerhalb des bereits erwähnten Energieversorgungsunternehmens erhoben wurden. Die einzelnen Erhebungsinstrumente, Messmethoden und Stichproben werden in den einzelnen Beiträgen jeweils herausgearbeitet.

Aufgrund der spezifischen Inhalte dieser Dissertation bietet sich eine Publikation in der Form von wissenschaftlichen Forschungsberichten (Vgl. Kapitel 3 bis bis 9) und nicht als traditionelle Monographie an. Dabei sind vier Forschungsberichte in Englisch (I, II, III, VII) verfasst und drei in Deutsch (IV, V; VI), welche durch weitere formale Bearbeitung auch in wissenschaftlichen Zeitschriften publiziert werden können. Die sieben Beiträge sind grundsätzlich eigenständig und können somit ohne spezifisches Wissen aus den anderen Kapiteln / Forschungsberichten gelesen und verstanden werden. Dies führt zu einigen redundanten Abschnitten, vor allem in den theoretischen Teilen, da jeweils ein Überblick des aktuellen Forschungsgegenstandes der jeweiligen Thematik gegeben wird.

Trotz ihrer Eigenständigkeit bilden die sieben Beiträge jedoch auch ein zusammenhängendes Ganzes rund um die drei zentralen Forschungsfragen. Einige Kapitel gehen derselben Frage mit unterschiedlichen Untersuchungsmethoden nach oder setzen unterschiedliche Schwerpunkte. In den folgenden Kapiteln wird nach einer kurzen theoretischen Einführung die wesentlichen Fragestellungen, die inhaltlichen und methodischen Ansätze sowie die Hauptbefunde der jeweiligen Forschungsbeiträge kurz umrissen.

## 2.1 Beurteilung der Qualität verschiedener Skalierungsverfahren

Für die Personalbeurteilung stehen verschiedene Messmethoden zur Verfügung. Personalbeurteilung kann mittels freier Eindrucksschilderung oder verschiedener Skalierungsverfahren (Auswahl-, Rangordnung- oder Einstufungsverfahren) vorgenommen werden. Dabei wird zwischen Selbst- und Fremdbeurteilung unterschieden. Zudem können Beurteilungsinstrumente entweder summarisch (d.h. ganzheitlich) oder analytisch (d.h. auf einzelnen Merkmalen) ausgelegt werden (Klimecki & Gmür, 1998). Auf inhaltlicher Ebene können prinzipiell Eigenschaften, Fähigkeiten, Verhalten und/oder Ergebnisse betrachtet werden (für einen Überblick: Liebel & Oechsler, 1994; Schuler, 2006).

In dieser Arbeit werden Rangordnungs- und Einstufungsverfahren miteinander verglichen und einander gegenübergestellt. Im dritten Forschungsbeitrag wird auch die summarische versus die analytische Beurteilung miteinander verglichen.

Während Rangordnungsverfahren mit einem Forced-Choice-Format operieren und Beurteiler „zwingen“, eine Rangreihe von Personen oder eine von Merkmalen von Personen zu bilden, entsprechen die Einstufungsverfahren den normativen Verfahren, bei denen die Merkmale von Personen (z.B. Kompetenzdimensionen) auf einer mehrstufigen Skala eingeschätzt werden und mit anderen Personen (zumeist einer Normstichprobe) in Bezug gesetzt werden (Saville & Wilson, 1991).

Beim Forced-Choice-Format sollen die Beurteiler ein ganz bestimmtes Kompetenzprofil einer Person zeichnen, indem die Stärken und Schwächen eines Mitarbeiters in Form einer Rangreihe von Kompetenzbegriffen eingestuft werden. Es wird dabei nicht die absolute Höhe der Kompetenzausprägung gemessen, sondern nur ein relatives Profil basierend auf den Kompetenzen erstellt. Für den interindividuellen Leistungsvergleich kann demnach keine Aussage über die Höhe dieser Profile gemacht werden. Solche auf Forced-Choice-Formaten beruhende Verfahren werden in der Literatur als „ipsative“ Verfahren bezeichnet (Saville & Wilson, 1991). Bei ipsativen Verfahren muss für die interindividuelle Vergleichbarkeit die Messung der absoluten Kompetenzhöhe bzw. der Gesamtleistung (Summarisches Urteil über die Kompetenz- oder Leistungshöhe) in einem separaten Schritt erhoben werden, indem nach der absoluten Kompetenzhöhe gefragt wird. So könnte man zum Beispiel den Beurteiler bitten, eine Rangreihe der leistungsstärksten Mitarbeiter zu bilden.

Den Forced-Choice-Verfahren stehen die gängigen Fragebögen gegenüber, welche häufig in Leistungsbeurteilungsbögen, Persönlichkeitsinventare oder in einem Kompetenzfragebogen verwendet werden und zu den normativen Verfahren zählen. Die Beurteiler kreuzen dabei auf einer intervallskalierten Mehrpunkt-Skala einen bestimmten Wert an, je nach dem wie sie die Ausprägung auf einem bestimmten Item einschätzen. Die Summe der angekreuzten Items ergibt dann den Ausprägungsgrad einer Dimension, welche jeweils mit der Normstichprobe in Bezug gesetzt wird. Bei dem hier beschriebenen Fragebogen-Formaten müssen die Beurteiler bei der Bewertung von Personen sowohl die Profilinformation über die verschiedenen Kompetenzdimensionen als auch die absolute Profilhöhe in Bezug zu anderen Personen im Auge behalten.

Ein wesentliches Forschungsinteresse der vorliegenden Arbeit liegt somit im Vergleich der beiden Verfahren in Bezug auf ihre Messgenauigkeit. Der direkte Vergleich von ipsativer- und normativer Kompetenzmessung wurde bislang eher wenig geforscht. Ein gewisser Schwerpunkt des diesbezüglichen Forschungsinteresses bildete sich vor allem in den 1990er Jahren heraus, als Saville & Willson (1991) sich um die Validität ihres bekannten Occupational Personality Inventory (OPQ32), eines der wenigen Forced-Choice-Messinstrumenten auf dem Markt, kümmerten. Ein Versuch einer Renaissance wurde durch den Artikel von Christiansen, Burns & Montgomery (2005) mit ihrem Artikel *Reconsidering Forced-Choice Item Formats for Applicant Personality Assessment* lanciert. Obwohl hohe Konstrukt- und prognostische Validität von Forced-Choice-Verfahren bestätigt werden konnten, beruhen die Ergebnisse einerseits lediglich auf Selbstbeurteilungsdaten und andererseits wurden nur Stichproben von Studenten erhoben (und keine jobrelevanten Daten aus der Praxis), was übrigens auch von den Autoren selbst bemängelt wurde.

Die in dieser Dissertation gefundenen Erkenntnisse über die Möglichkeiten verschiedener Messverfahren beruhen zum Teil auf „scharfen“ Daten, welche von Führungskräften der Energiewirtschaft gewonnen werden konnten. Dabei wurden zudem mittels online- und computergestützten face-to-face Befragungen drei unterschiedliche Datensätze aus drei verschiedenen Unternehmen der Schweizerischen Energiewirtschaft erhoben. In den verschiedenen Studien wurden sowohl die Antwortformate variiert als auch Fremdbeurteilungs- mit Selbstbeurteilungsdaten verglichen.

Die ersten beiden Forschungsbeiträge fokussieren im Hinblick auf die Qualität verschiedener Skalierungsverfahren auf die Beurteilerübereinstimmung multiperspektivischer Leistungsbeurteilung. Es mag nicht erstaunen, dass die Beurteilerübereinstimmung zwischen verschiedenen Beurteilern üblicherweise nicht sonderlich hoch ist. So konstatieren Facteau & Craig (2001) in ihrem Artikel: „One of the most constistent findings in the empirical literature on performance appraisal systems is that the ratings obtained from different sources generally do not converge“.

Die Frage stellt sich nun, worauf die unterschiedliche Beurteilung am ehesten zurückzuführen ist. Mögliche Quellen von Varianzen sind die tatsächlichen Leistungsunterschiede der Beurteilten, die verschiedenen Massstäbe der Beurteiler oder etwa das angewendete Messinstrument.

Die zweite Frage, die sich dran anschliesst, ist, ob je nach Anwendung des Messinstruments die Beurteilerinkongruenzen geringer werden. Facteau & Craig (2001) haben mit konfirmatorischer Faktorenanalyse getestet, ob verschie-dene Beurteiler Gruppen ein gemeinsames Verständnis von Leistung in Bezug auf eine bestimmte Faktorenstruktur haben. Das heisst, ihr Modell testet, ob die Struktur (d.h. die Anzahl latenter Variablen und die Items die mit den latenten Variablen verbunden sind) hinter den Beurteilern über die verschiedenen Beurteilergruppen „invariant", also unabhängig ist. Falls diese Unabhängigkeit bestätigt werden sollte, so würde dies laut Facteau und Craig (2001) bedeuten, dass die Inkongruenzen nicht durch das Messinstrument bedingt sind. Diese Unabhängigkeit des Messinstruments konnte in ihrer Studie dann auch bestätigt werden. An diesem Ergebnis knüpft die hier vorliegende Arbeit an und überprüft, inwiefern die Beurteilerinkongruenzen tatsächlich vom Messinstrument unabhängig sind. Dazu sollen zwei völlig verschiedene Messverfahren basierend auf denselben latenten Variablen miteinander verglichen werden. Die Abhängigkeit des Messinstruments wird dabei nicht mit konfirmatorischer Faktorenanalyse, sondern mittels Nonmetrischer Multidimensionaler Skalierung (NMDS) untersucht.

Basierend auf einem Datensatz, welcher im Rahmen einer Kaderbeurteilung eines grösseren zentralschweizerischen Energieunternehmens erhoben wurde, wurden Daten von 15 Führungskräften erfragt, welche von 5 verschiedenen Beurteilern mittels zwei unterschiedlichen Messinstrumenten (einem gängigen Kompetenzfragebogen und einem kompetenzbasierten Forced-Choice-Verfahren) eingeschätzt wurden. Das Hauptinteresse der Erhebung lag darin, die qualitativen Unterschiede der beiden Erhebungsinstrumente genauer zu eruieren und zu verstehen. Dabei wird der Kongruenz der verschiedenen Beurteiler in Abhängigkeit des angewendeten Messverfahrens besondere Beachtung geschenkt. Diese Studien werden in den beiden ersten Forschungsbeiträgen berichtet.

**Forschungsbeitrag I (Kapitel 3)**

Die erste Studie fokussiert auf das Antwortformat eines gängigen Fragebogens, bei welchem die Beurteilten auf mehreren Dimensionen auf einer Mehrpunktskala eingeschätzt werden. Es soll gezeigt werden, wie gängige Fragebögen starken Beurteilertendenzen unterliegen, da aufgrund des absoluten Formats die unterschiedlichen Massstäbe der Beurteiler stark ins Gewicht fallen. Wir vermuten

dabei einen starken Halo-Effekt durch einen generellen Leistungsfaktor, welcher die dimensionale Differenzierung auf einzelnen Kompetenzen überstrahlt. Mittels NMDS wird die Beurteilerübereinstimmung auf holistischer und dimensionaler Ebene betrachtet und miteinander verglichen. Als Hauptbefund wird sich zeigen, dass Inkongruenzen zwischen den Beurteilern vor allem durch den Beurteiler und dessen Anwendung eines unterschiedlichen Massstabs verursacht wird. Zudem wird gezeigt, dass sich eine grössere Übereinstimmung zwischen Beurteilern auf holistischer gegenüber dimensionaler Ebene bestätigt. Die Ursachen für die grössere Diskrepanz auf dimensionaler Ebene werden im Diskussionsteil des ersten Beitrages erörtert.

**Forschungsbeitrag II (Kapitel 4)**

Der zweite Forschungsbeitrag baut auf der ersten Fragestellung auf und stellt die Ergebnisse der Daten aus den Fragebögen mit den Daten aus dem Forced-Choice-Antwortformat gegenüber, bei dem die Versuchspersonen Kompetenzprofile in Form von Rangreihen produzieren. Es interessiert primär die Frage, inwiefern die Kongruenz der Beurteilten vom jeweiligen Messinstrument abhängig ist. Es wird die Hypothese aufgestellt, dass der Beurteiler-Effekt vor allem beim Fragebogenformat zum Tragen kommt. Basierend auf NMDS wird mittels Vergleich der Intraclusterdistanzen zwischen Beurteilern und Beurteilten gezeigt werden, dass ein Forced-Choice-Format eine höhere Beurteiltenübereinstimmung erreicht als das gängige Fragebogenformat. Der starke Halo-Effekt, welcher hauptsächlich auf dem unterschiedlichen Massstab der Beurteiler beruht, wird durch das Forced-Choice-Format abgeschwächt. Auf dieser Erkenntnis leitet sich die Empfehlung ab, Personalbeurteilungen in zwei Schritten durchzuführen: In einem ersten Schritt die kompetenzbasierte Profilbildung mittels einem Forced-Choice-Format und einem zweiten Schritt die Messung der absoluten Profilhöhe, also der Gesamtleistung bzw. Kompetenzhöhe im Vergleich zu anderen Personen. Dieses Vorgehen und die daraus gewonnenen Daten werden im dritten Forschungsbeitrag untersucht.

Auf einem zweiten Datensatz beruht der dritte Forschungsbeitrag, welcher in einem anderen grösseren Energieunternehmen im Rahmen der jährlichen Leistungsbeurteilung von Mitarbeitern erhoben wurde. Es handelt sich dabei um eine Längsschnittstudie, bei der Daten über ein Jahr zu drei verschiedenen Zeitpunkten erhoben wurden.

**Forschungsbeitrag III (Kapitel 5)**

Der dritte Forschungsbeitrag fokussiert auf die Fragestellung, ob neben der Profilinformation auch die absolute Profilhöhe durch ein Forced-Choice-Verfahren erhoben werden kann. Dabei wird zum einen ein Fragebogen mit mehreren Verhaltensdimensionen herangezogen, welcher im Rahmen der jährlichen Leistungsbeurteilung eines schweizerischen Energieunternehmens ohnehin verwendet wird, zum anderen wird dieselbe Stichprobe von den Beurteilern mit einem Forced-Choice-Rangordnungsverfahren konfrontiert, bei dem der Vorgesetzte seine Mitarbeitenden nach ihrer Gesamtleistung in eine Rangreihe bringt.

Es stellt sich natürlich die Frage, inwiefern ein analytisches Leistungsbeurteilungsverfahren in Bezug auf den Gesamtsummenscore zu denselben Ergebnissen führt wie das summarische Forced-Choice-Verfahren. Die Hypothese lautet, dass das summarische Rangordnungsverfahren (Forced-Choice) dieselbe prognostische Validität bietet wie das analytische Verfahren. Die Ergebnisse werden zeigen, dass sich zwischen dem summarischen Verfahren und dem analytischen Verfahren einen hochsignifikanten Zusammenhang ergibt. Dieser Befund unterstützt die Anwendung von Forced-Choice-Verfahren als Verfahren der Wahl sowohl für die Profilbildung als auch für die Beurteilung des ganzheitlichen Leistungs- bzw. Kompetenzniveaus von Mitarbeitenden.

Neben dieser kurz skizzierten Fragestellung interessiert weiter, welche Kompetenzen die Gesamtleistung am besten prognostizieren. Mittels Regressionsanalyse soll gezeigt werden, welche Kompetenzklasse (fachliche, methodische, soziale, persönliche Kompetenzen) die Gesamtleistung am besten prognostizieren.

Entgegen den zahlreichen Studien zur Bedeutung der sozialen Kompetenz oder emotionalen Intelligenz sind es vielmehr die fachlichen- und methodischen Kompetenzen als die sozialen und persönlichen Kompetenzen, die die Gesamtleistung der Mitarbeitenden am besten voraussagen. Inwiefern dieser Befund mit der „technikorientierten Stichprobe“ eines Energieunternehmens zusammenhängt, wird im Diskussionsteil abschliessend erörtert.

In den Forschungsbeiträgen IV und V beschäftigen wir uns mit der Konstruktion eines Selbstbeurteilungsinstruments, welches basierend auf den Erkenntnissen der ersten drei Studien auf einem Forced-Choice-Verfahren beruht. Anhand eines dritten Datensatz mit einer neuen Stichprobe von Führungskräften aus mehreren

Schweizerischen Energieunternehmen wurde das Kompetenzmessverfahren mittels NMDS geprüft.

**Forschungsbeitrag IV (Kapitel 6)**

Der vierte Forschungsbeitrag stellt die Konstruktion des kompetenzbasierten Selbstbeurteilungsverfahrens in den Vordergrund. Es wird dabei der Frage nachgegangen, inwiefern sich ein Forced-Choice-Verfahren auch für Selbstbeurteilungen eignet. Im Gegensatz zu den meisten Persönlichkeitsfragebögen, welche Selbstbeurteilungsdaten in Analogie zu Fremdbeurteilungsbögen auf mehrstufigen Skalen erheben und den summierten Werte einer Normstichprobe gegenüberstellen, wird hier ein gänzlich anderes Verfahren vorgestellt: Basierend auf den Erkenntnissen der ersten drei Studien, wird ein einfaches und stufenweise funktionierendes Selbstbeurteilungsinstrument entwickelt. Zur Entwicklung dieses Instruments werden die Beurteiler gebeten, anhand einer breiten Fülle von fachübergreifenden Kompetenzitems in einem mehrstufigen Forced-Choice-Verfahren eine Selbsteinschätzung ihrer Stärken vorzunehmen. Diese auf Rangdaten basierten Kompetenzprofile lassen sich mit Hilfe der NMDS in einem zweidimensionalen Raum als relationale Struktur abbilden.

Die deskriptive Validierung der NMDS Karten erfolgte durch die Anwendung eines auf langjährigen empirischen Untersuchungen beruhenden Kompetenzmodells von Wunderer (2001), welches die drei Dimensionen Gestaltungskompetenz, soziale Kompetenz und Umsetzungskompetenz beinhaltet. Mittels Expertenurteil wurden alle dem Forced-Choice-Verfahren zugrunde liegenden Komptenzitems in die drei Kompetenzkategorien eingeordnet und mittels Property fitting in die NMDS Karte gelegt. Als Hauptbefund konnte gezeigt werden, dass die drei Dimensionen von Wunderer sich sehr gut eignen, um die erhobenen Kompetenzprofile in der zweidimensionalen Karte sinnvoll und einfach zu interpretieren.

**Forschungsbeitrag V (Kapitel 7)**

Der fünfte Forschungsbeitrag ist eine Erweiterung des vierten Beitrages und hebt die Vorteile eines Forced-Choice-Verfahrens in den Vordergrund.

Es wird gezeigt, dass ipsative Messung in Form eines Forced-Choice-Verfahrens als eine echte Alternative zu den gängigen Ratingskalen normativer Messung betrachtet werden kann. Zum anderen kann als weiterer Hauptbefund aufgeführt werden, dass

die Messung von Kompetenzprofilen mittels eines Forced-Choice-Ansatzes relativ unabhängig von der Wahl bestimmter Kompetenzitems beziehungsweise deren spezifischen Terminologie ist. Mit dem fünften Forschungsbeitrag werden zudem die methodischen Vorteile der NMDS gegenüber der Faktorenanalyse aufgezeigt, indem die Kompetenzitems nicht nur mittels Faktorenanalyse, sondern auch mittels NMDS ausgewertet werden. Anhand NDMS Karten wird gezeigt, dass die semantischen Übergänge zwischen verschiedenen Kompetenzitems graduelle Übergänge darstellen, welche in faktoranalytischen Auswertungen verborgen bleiben.

## 2.2 Übereinstimmung zwischen Selbst- und Fremdbeurteilung

Die zweite Hauptfrage der vorliegenden Arbeit behandelt die Übereinstimmung zwischen Selbst- und Fremdbeurteilung.

Mit Abstand am häufigsten werden Fragebogen zur Messung von Führungsverhalten eingesetzt. Der wohl gewichtigste Vorteil dabei ist, dass auf ökonomische Art und Weise Fremd- und Selbstbeschreibungen kombiniert werden können. Dazu muss der Fragebogen lediglich von der Führungskraft selbst und zusätzlich von weiteren Personen bearbeitet werden. Was aber findet man vor in Unternehmen, wenn Selbst- und Fremdbilder miteinander verglichen werden? Man kann sagen: viele Wirklichkeiten – je nach Perspektive und individuellen Motiven fällt die Bewertung des Leistungsverhaltens einer Person anders aus.

Verschiedene Meta-Analysen haben das Ausmass der Diskrepanz zwischen Selbst- und Fremdbeurteilung untersucht. Auf korrelativer Ebene finden sich nur geringe Zusammenhänge zwischen dem Selbstbild der Führungskraft und den Fremdbildern. Für allgemeine berufliche Leistungsbeurteilungen schätzen verschiedene Meta-Analysen Zusammenhänge zwischen 0.19 und .35 (Vgl. Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988; Heidemeier & Moser, 2002 und 2009; Fleenor, McCauley und Brutus,1996).

Besonders interessant und Ausgangspunkt für die in dieser Arbeit relevante Fragestellung ist die Tatsache, dass die korrelativen Zusammenhänge steigen, wenn Fremdbilder miteinander verglichen werden und das Selbstbild aussen vor bleibt. Diesen Befund erbringen sowohl die Meta-Analysen für die Führungsbeurteilung von Führungskräften (Conway & Huffcutt, 1997/Harris & Schaubroeck, 1988: Aufwärts-Abwärts: $\rho$ = .57/-, Aufwärts-Peer: $\rho$ = .66/-, Abwärts-Peer: $\rho$ = .79/.62) als auch Einzelstudien (Fleenor et al. (1996): $r$ = .41 für Aufwärts-Abwärts- Korrelation bzw.

*r* = .24/.19 für Selbst-Aufwärts/Abwärts-Korrelation; Furnham & Stringfield (1998): *r* = .58 als mittlere Korrelation aus 72 Fremd-Fremd-Korrelationen bzw. *r* = .13 aus Selbst-Fremd-Korrelationen.

Generell wurde in der Forschung deutlich mehr Wert auf korrelative Zusammenhänge zwischen den Perspektiven als auf Niveau-Unterschiede gelegt. In den Forschungsbeiträgen I und II wurde bereits auf die starken Niveau-Unterschiede zwischen verschiedenen Beurteilern insbesondere bei normativen Verfahren hingewiesen.

**Forschungsbericht VI**

Der sechste Forschungsbericht greift die oben zitierte Befundlage auf, verändert jedoch in Bezug auf diese mehrheitlich aus dem amerikanischen Raum stammenden Studien, zwei wesentliche Parameter: Erstens wurden die erhobenen Daten nicht mittels den gängigen Likert-Skala basierten Fragebogenformaten, sondern mittels Forced-Choice-Format erhoben und zweitens wurde die Übereinstimmung zwischen Fremdurteilen und Selbsturteilen bzw. zwischen Fremdurteilen untereinander nicht mittels gemittelten Korrelationskoeffizienten bestimmt, sondern mittels NMDS unter Verwendung von Pearson-Korrelationen.

Basierend auf der extrahierten Distanzmatrix der NMDS-Karte wurde ein kritisches Distanzmass bestimmt, dass die Wahrnehmungskongruenz bzw. -Inkongruenz verschiedener Beurteiler zu einer Person beschreibt. Es kann gezeigt werden, dass die Fremdbilder zu einer Person sich insgesamt ähnlicher sind als die Fremdbilder über alle Beurteilten. Dieser Effekt kippt in die andere Richtung wenn man die Selbst- und Fremdbilder miteinander vergleicht.

Diese simultane Überprüfung von Fremd-Fremdurteilen und Selbst-Fremdurteilen ist ein zentraler Vorteil der NMDS, sobaldsowohl Beurteiler als auch Beurteilte in einem gemeinsamen zweidimensionalen Raum skaliert und auf einen Blick visualisiert werden. Durch diese Methode wird ersichtlich, inwiefern Selbst- und Fremdurteile tatsächlich zwei „unterschiedliche Welten“ darstellen.

## 2.3 Die Überprüfung der Reliabilität eines bevorzugten Skalierungsverfahrens

Der Abschluss dieser Arbeit bildet die Frage der Reliabilität von Selbstbeurteilungsdaten. Es stellt sich die Frage, wie stabil Selbstbeurteilungen über die Zeit hinweg sind. Bisher veröffentlichte Studien berichten von mittleren Retest-Reliabilität von bis zu $r$ = .81 (Nilsen & Campbell, 1993; Smither, London, Vasilopoulos, Reilly, Millsap & Salvemini, 1995). Salgado et al. (2003) bemängeln eine geringe Anzahl an Studien zur Retest-Reliabilität. Mir persönlich sind keine Forschungsbeiträge zu Retest-Reliabilitäten von Force-Choice-Formaten mit echten Daten aus der Praxis bekannt.
Der siebte und letzte Forschungsbeitrag dieser Arbeit untersucht deswegen die Retest-Reliabilität von Forced-Choice-Daten anhand der NDMS Methode.

**Forschungsbeitrag VII**

Basierend auf den Daten der Forschungsbeiträge IV bis VI, wird im siebten Forschungsbeitrag der Fokus auf die Stabilität der kompetenzbasierten Forced-Choice-Profile gelegt. Im Vordergrund steht dabei die Frage, inwiefern die zeitliche Stabilität von Selbstbeurteilungsdaten vom Antwortformat und vom Itemformat abhängig ist.
Die Hypothese lautet, dass ein Selbstbeurteilungsverfahren basierend auf einem Forced-Choice-Format mit Stabilitätskoeffizienten von gängigen Selbstbeurteilungsinventaren wie zum Beispiel dem Eysenck Personality Inventory (Eysenck & Eysenck, 1975) oder dem Bochumer Inventar zur berufbezogenen Persönlichkeitsbeschreibung (Hossiep & Paschen, 2003) durchaus mithalten kann. Zudem wird der Frage nachgegangen, inwiefern das Itemformat einen Einfluss auf die Stabilität von Selbstbeurteilungsdaten hat. Es wird zwischen situationsspezifischen Verhaltensankern und situationsübergreifenden Kompetenzbegriffen unterschieden, welche je in einer der zwei Messbedingungen erhoben wurden. Methodisch wird die Überprüfung der Stabilität sowohl des Antwort- als auch des Itemformats mittels NMDS durchgeführt, wobei ein Testmodell zur Überprüfung der Retest-Reliabilität vorgestellt wird, das die Ansprüche der gängigen Test-Retest-Korrelationskoeffizienten übertrifft.

## 2.4 Literatur

Atwater, L.E., Ostroff, C., Yammarino, F.J., & Fleenor, J.W. (1998). Self-other agreement: does it really matter*? Personnel Psychology*, 51, 577-598.

Atwater, L.E., & Yammarino, F.J. (1992). Does self-other agreement on leadership perceptions moderate the validity of leadership and performance predictions? *Personnel Psychology*, 45, 141-164.

Baril, G. L., Ayman, R., & Palmiter, D. J. (1994). Measuring leader behavior: Moderators of discrepant self and subordinate descriptions. *Journal of Applied Social Psychology*, 24(1), 82-94.

Barrett, G.V. & Depinet, R.L. (1991). A reconsideration of testing for competence rather than for intelligence. *American Psychologist*, 46, 1012-1024.

Berthel, J. (2000). Personalmanagement, 6. Auflage, Stuttgart.

Borg, I., Groenen, P & Mair, P. (2010). Multidimensionale Skalierung.In: *Sozialwissenschaftliche Forschungsmethoden*, Band 1, Rainer Hampp Verlag, München u. Mering

Christiansen, N.D., Burns, G.N. & Montgomery, G.E, (2005). Reconsidering Forced-Choice item formats for applicant personality assessment. *Human Performance*, 18, 267-307.

Church, A. H.(1997a). Managerial self-awareness in high-performing individuals in organizations. *Journalof Applied Psychology, 83*, 281-292.

Conway, J. M. &Huffcutt, A. I. (1997). Psychometric properties of multisource performanceratings: Ameta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance*, *10*, 331-360.

Eysenck, H. J. & Eysenck, S. B. G. (1975). Manual of the Eysenck Personality Questionnaire. San Diego: Educational and Industrial Testing Service.

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology*, 86, 215 227.

Ferris, G.H (2006). Research in Personnel and Human Resources Management, 19, Greenwich, CT: JAI Press.

Fersch, J.M. (2002). Leistungsbeurteilung und Zielvereinbarungen in Unternehmen,1. Auflage, Wiesbaden.

Fleenor J, McCauley C, Brutus S. (1996). Self-other rating agreement and leader effectiveness. *Leadership Quarterly*, 7, 487-506.

Furnham, A. and Stringfield, P. (1998), "Congruence in job-performance ratings", *Human Relations*, 51, 517-30.

Harris, M. M. &Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisorratings. *Personnel Psychology, 41*, 43-62.

Heidemeier, H.& Moser, K. (2002). *Self-appraisal of job-performance*. Poster beim 43. Kongress derDeutschen Gesellschaft für Psychologie.

Heidemeier, H. & Moser, K. (2009) „Self-other agreement in job performance ratings: A meta-analytical test of a process model“, *Journal of Applied Psychology*, 94, 353-370.

Hossiep, R. & Paschen, M. (2003, unter Mitarbeit von O. Mühlhaus). Bochumer In ventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP) (2. Aufl.). Göttingen: Hogrefe.

Klimecki, R. & Gmür, M (1998). Personalmanagement, 1. Auflage, Stuttgart, 1998.

Kurz, R. & Bartram, D. (2002). Competency and individual performance: Modelling the world of work. In I.T. Robertson, M. Callinan & D. Bartram (Eds.), *Organsational effectiveness: The role of psychology.* Chichester: Wiley.

Liebel, H.J. & Oechsle, W.A. (1992). Personalbeurteilung: Neue Wege der Leistungs- und Verhaltensbewertung, 2. Auflage, Bamber.

London, M. & Wohlers, A. J. (1991). Agreement between subordinate and self-ratings in upward feedback. *Personnel Psychology*, 44, 375-390.

Mabe, P. A. & West, S. G.(1982). Validityof self evaluations of ability: A review and metaanalysis. *Journalof Applied Psychology*, *67*, 280-296.

McClelland, D. C. (1973). Testing for competence rather than for "intelligence". *American Psychologist*, 28, 1-14.

Motowidlo, S. (2002). Job performance. In W. C. Borman, D. R. Ilgen, R. J. Klimoski, & I. B. Weiner (Eds.), *Handbook of Psychology, Industrial and Organizational Psychology*, 12, 39-53. New York: John Wiley and Sons.

Nilsen, D. and Campbell, D. (1993), ``Self-observer rating discrepancies: once an overrater, always an overrater?", *Human Resource Management*, 32, 265-281.

Salgado, J. F. (2003). Criterion validity of personality measures based and non-based on the Five Factor Model. *Journal of Occupational and Organizational Psychology*, 76, 373–385.

Sarges, W. (2006). Competencies statt Anforderungen – nur alter Wein in neuen Schläuchen? In H.-C. Riekhof (Hrsg.), *Strategien der Personalentwicklung* (6. Aufl.; S. 133-148). Wiesbaden: Gabler.

Saville, P., & Willson, E. (1991). The reliability and validity of normative and ipsa tive approaches in the measurement of personality. *Journal of Occupational Psychology*, 64(3), 219-238.

Schuler, H. (2006). Lehrbuch der Personalpsychologie, 2. Auflage, Göttingen: Hogrefe.

Smither, J.W., London, M., Vasilopoulos, N.L., Reilly, R.R., Millsap, R.E., & Salvemini, N. (1995). An examination of the effects of upward feedback over time. *Personnel Psychology*, 48, 1–34.

Viswesvaran, C. (2002) Assessment of individual job performance: A review of the past century and a look ahead. In Anderson, N. & Ones, D. S. (Eds.) *Handbook of industrial, work and organizational psychology,* Volume 1: Personnel Psychology. Sage Publications: London. 110-126.

Wunderer, R. & Dick, P. (2007). Personalmanagement - Quo vadis? Analysen und Prognosen zu Entwicklungstrends bis 2010: 5. aktualisierte Auflage. Köln: Luchterhand.

Zwell, M. (2000). Creating a culture of competence. New York: Wiley.

# 3 The quality of competency based performance ratings in a practical context

## 3.1 Abstract

The purpose of this study is to verify whether a multisource performance appraisal instrument applied in a practical context, exhibited measurement invariance across different types of raters under a dimensional and holistic point of view. The authors argue that conventional multipoint appraisal which tries to measure interindividual absolute performance level and intraindividual variance simultaneously lead to a cognitive bottleneck. They suggest disentangling profile information from absolute level information using a two step procedure of first assessing the overall performance level and secondly measure profile information through a forced-choice measure.

The implication of using competency based performance appraisal systems in order to increase quality of personnel judgement is critically assessed.

Nonmetric multidimensional scaling technique was used to demonstrate rater and ratee-caused effects in performance appraisal when applying traditional multipoint Likert-type items based on a leadership competency model. The results show, that a strong idiosyncratic rater effect occurs which is stronger than actual differences in performance attributed to ratees. In addition this study proves higher interrater reliability on a holistic rather than on a dimensional level which is explained with a general halo effect caused by focusing on an overall performance score.

## 3.2 Introduction

Job performance is a fundamentally important construct in organizational practice and research (Cleveland, Murphy, & Williams, 1989). From a practical perspective, it plays a central role in most personnel decisions, such as merit-based compensation, promotion, and retention. It is also used as an important source of developmental feedback and conceptualization of leadership programs. From a theoretical perspective, researchers have long been interested in understanding the causal mechanisms that lead to effective job performance.

Although job performance has been measured in many ways (e.g., volume of sales, quantity or quality of items produced, absences, number of promotions), the most

frequently used measure is a supervisory performance rating mainly based on a company wide used competency model. In recent years, multirater or 360°-feedback programs have gained popularity. This means that peer, subordinate, and self-ratings also play an important role in the assessment of job performance. Given the significance of the job performance construct and the growing dependence on ratings from multiple sources as a means for measuring that construct, it is important to identify to what extent different raters converge when judging a person's performance. Unfortunately, a number of studies conducted over the past several decades indicate that supervisory ratings often are plagued by a host of potential problems, including halo, leniency, intentional manipulation, and race, gender, or age biases (see Cardy & Dobbins, 1994; Cascio, 1991; Landy & Farr, 1980, for reviews).

Perhaps one of the most consistent findings in the empirical literature on performance appraisal systems is that the ratings obtained from different sources generally do not converge. The intercorrelations among the ratings provided by different raters tend to be moderate at best. For example, in a meta-analysis, Harris and Schaubroeck (1988) reported mean self-supervisor, self-peer, and peer-supervisor rating correlations (corrected for unreliability) of .35, .36, and .62, respectively. Similarly, Mount (1984) reported mean supervisor-subordinate and subordinate-self rating correlations of .24 and .19, respectively. Finally, Conway and Huffcutt (1997) reported that the correlations (corrected for unreliability) among ratings made by self, peer, supervisor, and subordinate raters ranged from a high of .79 (supervisor-peer) to a low of .14 (subordinate-self).

A number of different explanations have emerged to account for why different types of raters (e.g., peers, supervisors) do not agree in their ratings. Campbell and Lee (1988) argued that different rater groups may have different conceptualizations of what constitutes effective performance in a particular job. Murphy and Cleveland (1995) suggested that raters differ in their opportunity to observe any given individuals' work behavior and that these differences in perspective may account for disagreements among their ratings. Similarly, Lance, Teachout, and Donnelly (1992) noted that the modest correspondence among ratings from different types of raters may be due to the fact that they are exposed to only moderately overlapping sets of ratee behavior. This "ecological perspective" to rating source differences (Lance &

Woehr, 1989) suggests that strong correspondence among ratings from different sources should not be expected.

According to findings of Scullen, Mount and Goffrey (2000) it seems important to differentiate between convergence on a holistic or dimensional level when comparing scores of performance appraisal. Holistic performance is reflected in a general factor that underlies all judgments of a ratee's performance across all raters and performance dimensions (King, Hunter & Schmidt, 1980). This conceptualization of general performance is related to the concept of true halo (Cooper, 1981). Halo error refers to the tendency of raters to allow an overall impression of a ratee to influence judgements along several quasi-independent dimensions (King et al., 1980; Lance, LaPointe, & Stewart, 1994). Some degree of true halo is expected, because many of the antecedents of performance (e.g., mental ability and conscientiousness) are similar across the various dimensions of performance (Motowidlo, Borman, & Schmit, 1997). To the extent that ratings reflect actual performance, we expect to find evidence of a general factor in performance ratings. The second component of actual performance relates performance on a particular dimension to the ratee's general level of performance. Scullen, Mount and Goff (2000) found that the dimensional factors contributed fairly little unique information beyond what was associated with the general ratings factor. In other words, aspects of ratee performance that are specific to a particular dimension had a relatively minor influence on ratings. This finding is consistent with those of Viswesvaran (1993), who found support for a strong general performance factor in 25 distinct measures of job performance. It is not yet clear, however, why the general factor exerts a more powerful influence on ratings than do the dimensional factors.

The purpose of this study is to verify whether a multisource performance appraisal instrument applied in a practical context, exhibited measurement invariance across different types of raters under a dimensional and holistic point of view. The authors argue that conventional multipoint appraisal systems which aim to assess simultaneously the interindividual absolute performance level *and* the intraindividual variance on a dimensional level lead to a cognitive bottleneck. This assumption is connected with the "bottleneck" conception of human information processing (e.g., Broadbent, 1958; Treisman, 1969). This assumption can also be empirically examined when data is compared on the basis of Nonmetric Multidimensional Scaling (NMDS) using correlation coefficients and city block distances as proximity measures. The

assumption is, that on a correlation level, the convergence between different rater is higher than on the city block measure, since strong rater effect is mainly due to the different perception of absolute performance level. Based on the results, the authors would suggest disentangling profile information from absolute level information using a two step procedure of first assessing the interindividual overall performance level and secondly measure intraindividual profile information through a forced-choice measure. This is of main practical impact, since competency based performance ratings are, as initially mentioned, used for selection purposes (interindividual) as well as for developmental feedback and management development programs (intraindividual). But if we have large disagreement between different raters that is partly due to the measurement instrument, we might as well put some effort in further research using a different measurement instrument for performance appraisal. These important implications will be discussed later in this article.

## 3.3 Method

*Participants and design*

The study included a sample of 15 upper level managers (ratees) of a large utility company in Switzerland. The managers represented a variety of functional areas, such as power generation, power delivery, and customer operations. Because of concerns for anonymity, no information regarding raters' demographic characteristics was collected. The 15 managers were rated in a multisource rating design of 3 different functional roles: Chief Executive Officer (rater 1), 3 members of the executive board as direct line superiors of the departments “energy”, “grid” and “finance” (raters 2a, 2b, 2c) and Head Human Resources (rater 3). This is a typical setting which we encounter in practice. Rater 1 and 3 rated all ratees while raters 2a, 2b and 2c rated only the ratees in their department. This corresponds to a mixed design of cross and nested ratings. In cross rating systems, each rater rates the performance of all ratees. In nested ratings systems, each ratee’s performance is rated by different raters. This distinction is important because the amount of idiosyncratic ratings variance is expected to be larger in a nested design than in a crossed design. The reason is that each rater in a nested design may exhibit a different degree of leniency. Leniency errors refer to a rater’s tendency to assign ratings that are generally higher (or lower) than are warranted by the ratee’s actual performance. Thus, rater leniency differences

introduce an element of variability into nested designs that is not present in crossed designs. To the knowledge of the authors of the present study, the magnitude of the difference in idiosyncratic variance between these two different types of designs is not yet known, largely because research has failed to acknowledge the distinctions between the two types of designs. In this study, we assume that both, halo and leniency differences, contribute to idiosyncratic rating variance. Because the intention in this study was to quantify only the overall effect of rater bias, the distinction of leniency and halo effect was not attempted.

*Procedure*

The data for this study was collected during the administration of a competency based evaluation of the first management level (Top Management). The competency based questionnaire was identical for each rater group, and it contained 51 behaviourally oriented items that were designed to measure 17 dimensions of managerial competencies. According to Lievens, Sanchez & De Corte (2004) competency modelling ties the derivation of job specifications to the organization's strategy, which, together with non-strategic job requirements, are used to generate a "common language" in the form of a set of human attributes or individual competencies. The competence dimensions included entrepreneurship, customer orientation, goal and result orientation, innovation, planning & organizing, analysis and problem solving, implementation, initiative, reasoning and decision making, resilience, leadership, integrity, influence and control, conflict management, communication, teamwork and networking skills as well as professional knowledge. These competency dimensions were the result of a benchmark study[3] that compared researched based and practical proved competency models identified by the organization as being critical for successful leadership within the company. Managers who were being evaluated as part of the multisource competency evaluation completed self ratings of their performance. After rating forms were processed, managers received feedback about their ratings in the form of a feedback report and a feedback discussion based on the

[3] The benchmark study was based on competency research of Boyatzis (1982), Spencer and Spencer (1993), Thornton and Byham (1982), Shipman et al. (2000), the Swiss Military Leadership Institute, the Psychology Department of Applied Cognitive Sciences at University of Zürich and four different practical competency models used in different companies of the energy branch.

report. These reports summarized the ratings that managers received on each performance dimension as well as on each item by rating source. In this particular study the self-ratings were not considered.

*Measures*

Each of the 17 competency dimensions was described through 3 behavioural anchors. When completing the questionnaire, raters were encouraged to think about how often the manager they were rating demonstrated the behaviour described by each item. They were asked to base their ratings on behaviours that they had observed, not on how they believed a manager might behave. Ratings were made on an 8-point Likert-type scale that was segmented in 4 categories from “poor” to “excellent”. For each individual we obtained an overall competency score by summing up the score on all items as well as a competency profile representing the score aggregated for each dimension.

*Analysis*

The traditional approach when dealing with multitrait-multirater data is the application of confirmatory factor analysis. The correlated traits, correlated methods (CTCM, Widaman, 1985) model is often recommended (e.g., Kenny & Kashy, 1992), because it is most consistent with the Campbell and Fiske (1959) standards of psychometric measurement. However, the CTMC model was not a viable option in this study for both theoretical and practical reasons. Theoretically, the CTMC model is not completely consistent with the purpose of the present study. It also partitions variance into three components (trait, method, and error), but it does not simultaneously show the effect of general and dimensional performance in combination with rater bias and actual job performance differences.

In the present study we used the Nonmetric Multidimensional Scaling Technique (NMDS) in order to illustrate the latent structure of job performance ratings. The data of the competency scores on each of the 17 dimensions were copied into correlation matrices and then analyzed by robust NMDS (by means of the ROBUSCAL algorithm (Läge, 2001)). NMDS is a nonmetric algorithm, which interprets the competency scores on an ordinal level. In an iterative process, the configuration, which best corresponds to the proximities between the rated managers is approximated. The target configuration describes an n-dimensional space, which in

this study resulted in two dimensions. The NMDS depicts the cognitive relational structure of a subject in the form of an Euclidian space, where a small distance between the objects (i.e. competency profile of a manager) corresponds to a high competence profile similarity and vice versa. One may argue that in correlation measures the absolute competency level does not have any influence on the similarity between two objects. We therefore computed also NMDS taking into account the absolute score (i.e. overall competency score) using city block distances to calculate the matrices of competency profiles.

In order to interpret the map correctly we applied a multiple regression method known as property fitting (i.e. Läge, 2001) by putting an external criterion such as the overall competence score (sum of competence score over all 17 dimensions) into the map. In methodological terms this means that on the external scale, each object (ratee) is assigned a value with regard to the criterion that is hypothesized to underlie the NMDS configuration. If we assume for instance, that Manager K rated by rater 2a has a higher overall competence score than manager K rated by rater 1, then the former would receive a higher value on a "competence level" scale or dimension than the latter. These values are then fitted into the NMDS configuration by a multiple regression. A high multiple correlation of the dimension in the NMDS configuration indicates a high level of explanation.

Using NMDS we could scale all the 45 competency-profiles (15 ratees, 5 different raters) and examine the distance relations between the profiles. The goal of this paper is to visualize two major effects in competency based performance data:

- H1: a strong idiosyncratic rater effect explaining most of the variance in performance data
- H2: higher interrater reliability on a holistic than on a dimensional level

The first effect we illustrate using NMDS and hierarchical cluster analysis applying two different measures as explained above. We then compare intra- vs. interclusterdistances of the different objects, focusing on one hand on the ratees (distances between ratees) and on the other hand focusing on the different raters (distances between raters). This explains if the effect of rating differences are more due to the rater's different perception or the actual different competency levels of the ratees. In order to show the second effect we compare Pearson correlation coefficients

on a holistic (i.e summed up competence score) and on a dimensional level (i.e. each competency construct) over all rated managers.

## 3.4 Results

The first step in our analysis is to compare empirical data with an ideal data set of multiperspective performance appraisal. If we assume for instance, that all interrater correlations between tree raters over a sample of ratees are higher than r=.8, we would consider this as a sign for high congruence among raters. This would most likely mean that differences among ratees would only be accountable on the differences in performance and not be deteriorated by a rater or instrument bias. On a NMDS map this would be illustrated as the following.

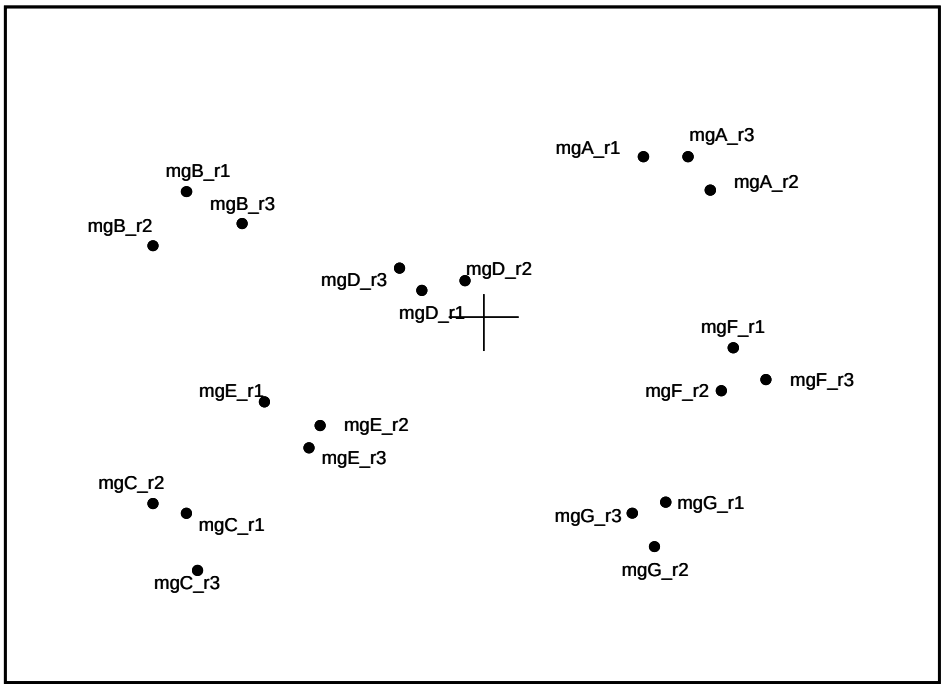
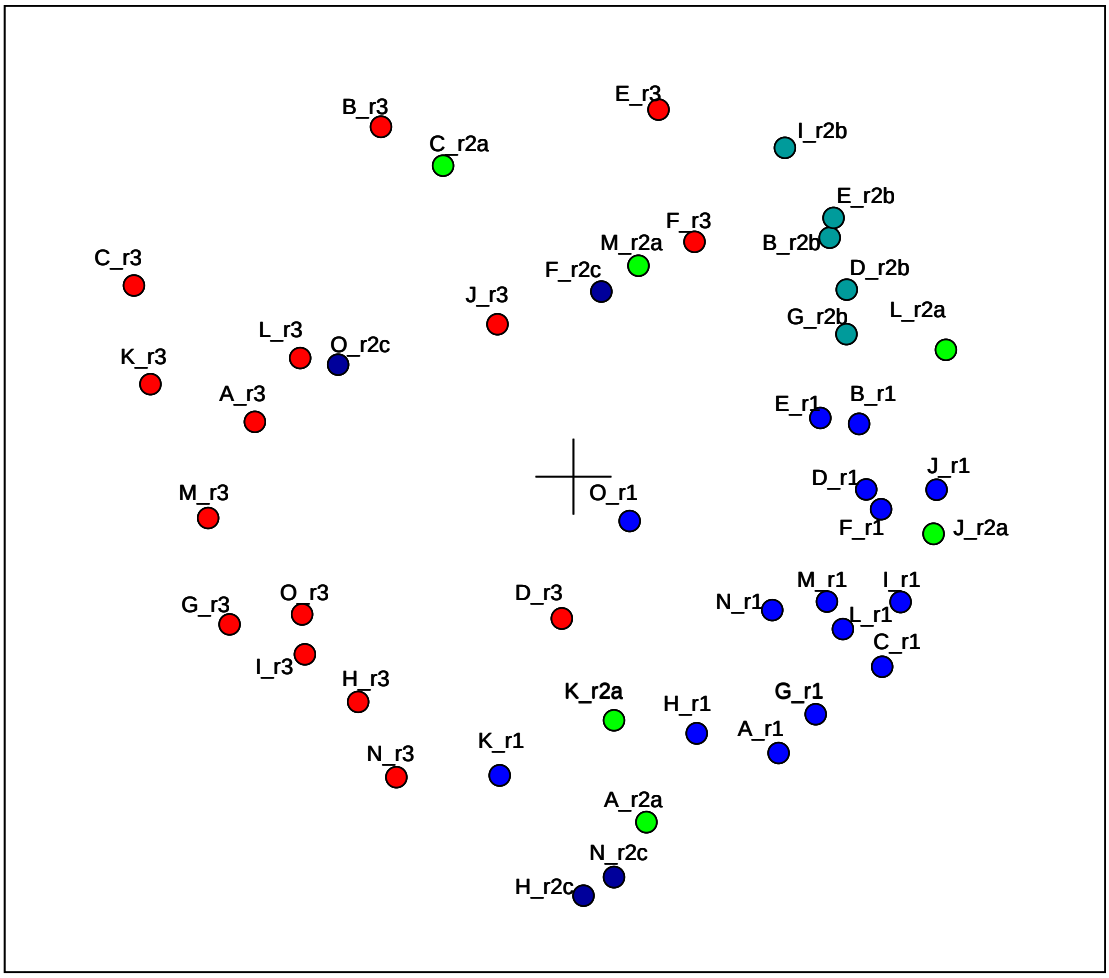



Figure 1: Hypothetical NMDS based on fictional ratings of 15 managers (mg=managers) rated by 3 different raters (r=rater)

In the map illustrated the clusters represent a unit of 3 different raters of one rated manager. The proximities between the points represent similarity between the competence profiles based on Pearson correlations.

If we look at the map generated by real performance data, we observe a rather unstructured set of points representing different managers. The distances again represent similarity of the competency profiles. Each manager is labelled with an alphabetical letter and indicated with the colour of the different raters. When comparing the ideal map to the empirical map based on real competency rating data,

we observe that interrater reliability is not as high as desired, since one manager rated by 5 different raters is not obviously represented in the same position in the NMDS map. Only managers E, O, F and H are somewhat in the same area of the map which indicates similar competency profiles. What seems far more obvious is the predominant influence of the rater.




Stress NMDS: 0.242

Figure 2: Relational positions of 45 object (competency profiles of 15 managers rated by 5 different raters) in a NMDS based on Pearson correlations.

In order to find a criterion to decide whether the ratings of a manager are consistent, or speaking in terms of a psychometric criterion, have a high interaterreliability, intra- and intercluster distances between managers were compared.

Intracluster distances represent the mean of the extracted distances of one manager rated by 3 different raters. Intercluster distances are the distances between the mean of one manager and the mean of distances to all other managers. Looking at the results on table 1.1 one can see that there is no systematic trend showing smaller intracluster

distances than intercluster distances. In 7 out of 15 cases we even found the inverse effect, with smaller inter- than intraclusterdistances.

| **managers** | **intracluster distances** | **intercluster distances** | **p-value** | **sign.** |
|---|---|---|---|---|
| *mg A* | *1.43* | *1.39* | *0.43* | *n.s* |
| mg B | 1.31 | 1.40 | 0.35 | n.s |
| *mg C* | *1.99* | *1.56* | *0.03* | *p<5%* |
| mg D | 1.04 | 1.18 | 0.18 | n.s |
| mg E | 0.81 | 1.37 | 0.02 | p<5% |
| mg F | 0.85 | 1.21 | 0.04 | p<5% |
| *mg G* | *1.78* | *1.34* | *0.00* | *p<0.1%* |
| mg H | 0.90 | 1.42 | 0.02 | p<5% |
| *mg I* | *1.90* | *1.38* | *0.00* | *P<0.1%* |
| mg J | 1.07 | 1.27 | 0.21 | n.s |
| *mg K* | *1.32* | *1.44* | *0.32* | *n.s* |
| *mg L* | *1.67* | *1.33* | *0.03* | *P<5%* |
| *mg M* | *1.62* | *1.29* | *0.00* | *P<0.1%* |
| mg N | 1.03 | 1.37 | 0.05 | n.s |
| Mg O | 0.99 | 1.28 | 0.00 | P<0.1% |
| *Mean* | **1.31** | **1.34** | **0.40** | **n.s** |

Table 1: Ratee effect: Intra- vs. interclass Distances of Pearson map

As stated earlier, what points out in the map is that the position of managers is highly dependent on the rater. When we focus on the positions of the ratees in dependency of the rater, we find clusters that are formed depending on the rater. This visual impression can be confirmed in the hierarchical cluster analysis.

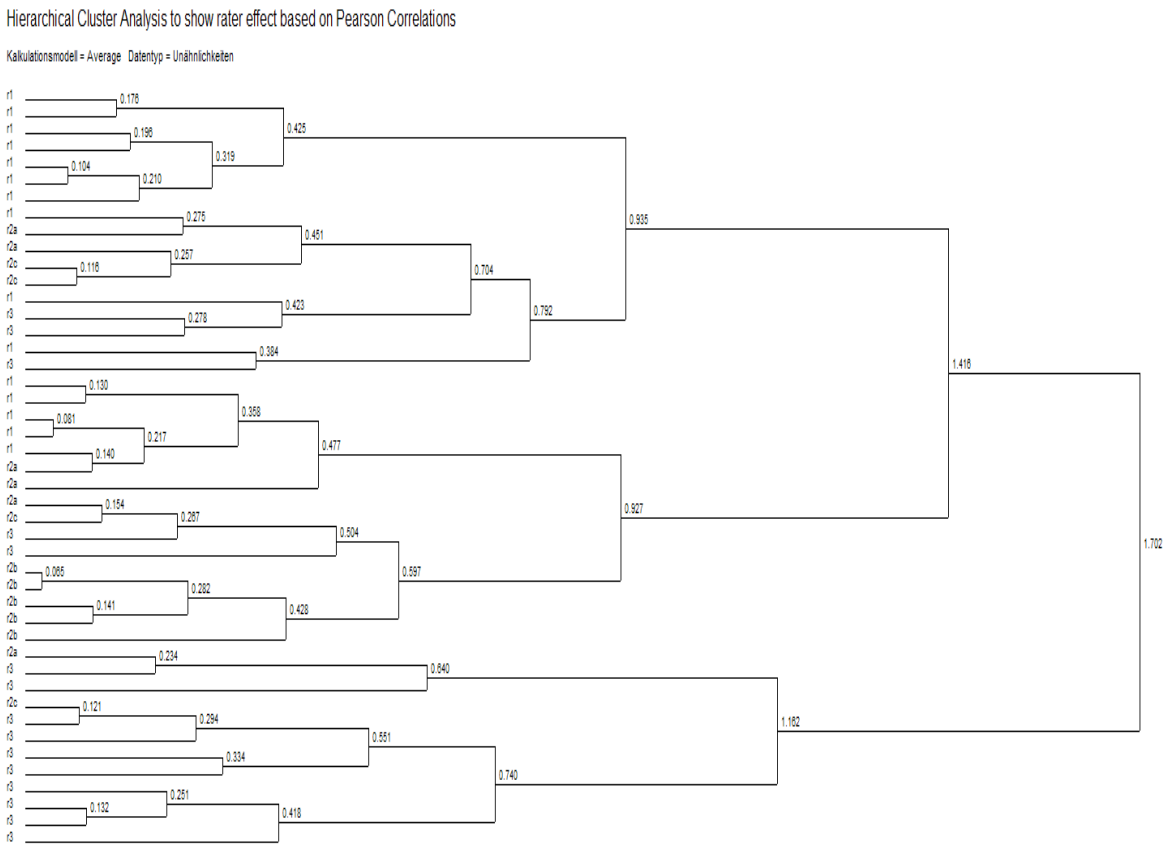



Figure 3: Hierarchical cluster analysis based on dissimilarities between 45 objects (15 managers rated by 5 different rating sources).

Cluster analysis confirms the assumption of a strong rater bias. Clusters are not built on similarity between ratees' competencies but on similarity of rating styles of the raters. We can therefore conclude that when focusing on the positions of the managers, one can argue that the ratings are relatively heterogeneous and interrater reliability is not given.

When considering the positions of the managers focusing on different raters, one can assume that the rater has a large effect on the ratings, since different managers rated by the same rater are positioned closely together in the map. An obvious exception of this assumption is rater "r2a" who's ratings are spread across the two-dimensional map. In order to proof this assumption, intra- and intracluster distances of raters have been compared.

| raters | intracluster | intercluster | p-value | sign. |
|---|---|---|---|---|
| **r1** | 0.66 | 1.33 | 1.09E-08 | p<0.01% |
| ***r2a*** | *1.40* | *1.30* | *0.053* | *n.s* |
| **r2b** | 0.30 | 1.46 | 1.90E-05 | p<0.01% |
| **r2c** | 1.43 | 1.45 | 0.42 | n.s |
| **r3** | 1.12 | 1.61 | 0.0003 | p<0.01% |

Table 2: Rater effect: Intra- vs. interclass Distances of Pearson map

Rater r1, r2b and r3 show significant smaller intracluster distances. This can be generally interpreted as a rather strong rater bias or as an overall rating style, which focuses on certain aspects when rating a person. From a content point of view the rater judgements on a person depend more on him than on the performance of the ratee. As explained in the methodological part of this research paper, one may correctly argue that correlations do not take into account the absolute level of ratings. High correlations may occur even though the level of ratings, and therefore the additive competence score, is significantly different. To show that a strong rater effect is also shown when taking into account the absolute level, city block distances had been used as a proximity measure to calculate the NMDS.

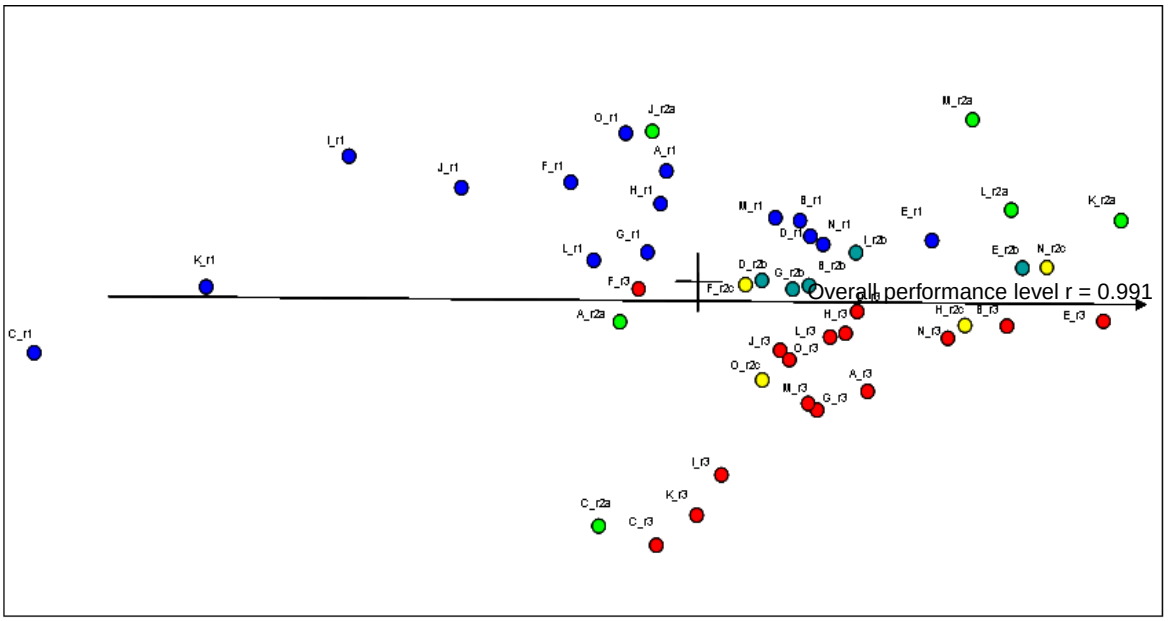



Figure 4: Relational positions of 45 object (competency profiles of 15 managers rated by 5 different raters) in a NMDS based on city block distances.

Looking at the map in figure 4 we can depict two remarkable effects. First of all it is quite noticeable how the absolute competence score of the ratings constitute the structure of the relational positions of the managers in a two dimensional Euclidian space. Looking at the high regression coefficient of 0.991 we can say with confidence that to the left on the map we find managers with low overall competence scores whereas on the right the more competent managers are represented. Second of all it is also conspicuous how strong rater effects can be shown.

Putting this picture into numbers, comparing again intra- and intercluster distances, the hypothesis of a strong rater effect can be partially confirmed. Rater 2b and rater 3 show strong rater effects with significant mean differences in intra- and intercluster distances. Considering the picture of the NMDS map one may suggest that there is also a strong rater bias regarding rater 1. Since manager C and K and I were rated extremely low and are therefore considered as outliers, the resulting distances become sufficiently large for voiding the effect of significant inter- and intracluster distances of managers rated by rater 1. However, the positions in the map are clear enough to see the idiosyncratic rating behaviour of rater 1 at one glance.

| **Rater** | **intracluster distances** | **intercluster distances** | **p-value** | **sign.** |
|---|---|---|---|---|
| **r1** | 1.72 | 1.82 | 0.28 | n.s |
| ***r2a*** | *1.75* | *1.59* | *0.15* | *n.s* |
| **r2b** | 0.59 | 1.11 | 3.21E-09 | p<0.01% |
| **r2c** | 1.01 | 1.29 | 0.06 | n.s |
| **r3** | 0.87 | 1.47 | 2.56E-05 | p<0.01% |

Table 3: Rater effect: Intra- vs. interclass distances of city block map

This relatively strong rater effect can also be shown by doing again a hierarchical cluster analysis based on dissimilarities.

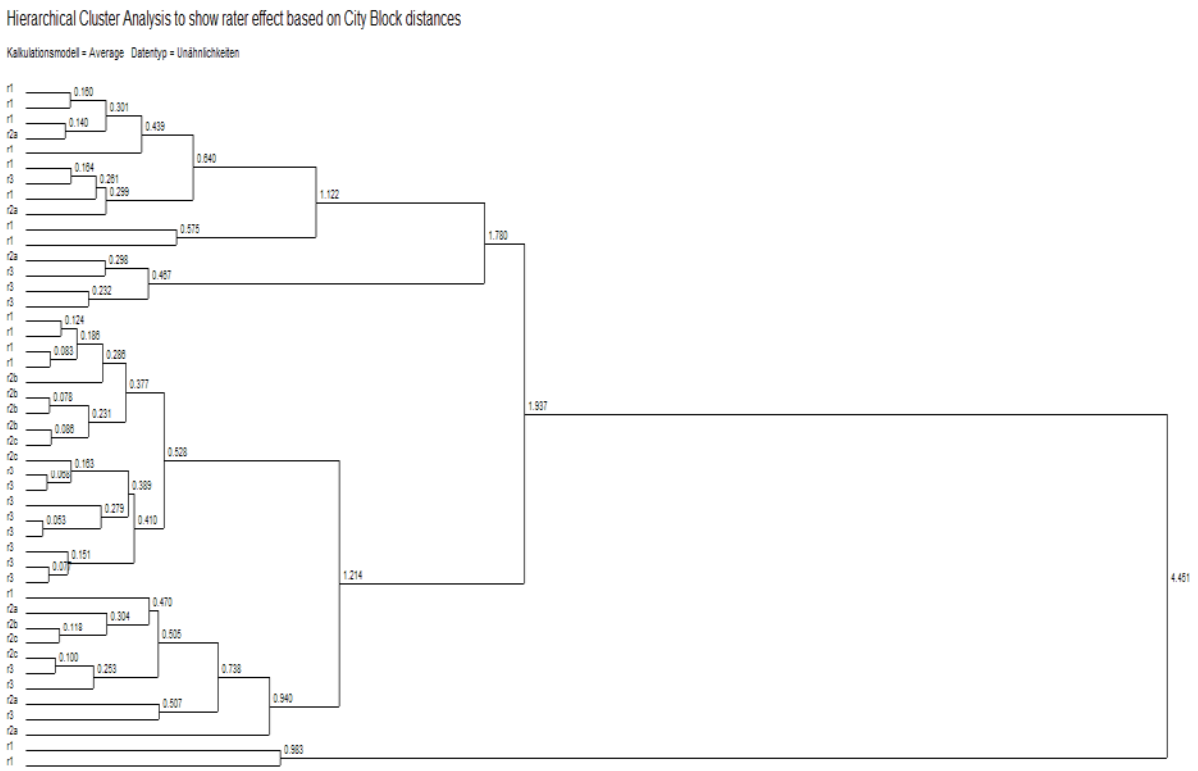



Figure 5: Hierarchical cluster analysis based on city block distances

Like in the hierarchical cluster analysis based on Pearson correlation we also find here a rater bias grouping clusters according to the different raters.

As in the NMDS based on Pearson correlation we can also compare the intra- and intercluster distances focusing on the ratees rather than the raters taking the absolute rating level into account. Even though on an individual level we do not find many significant differences between intra- and intercluster distances, when comparing on a mean level, we depict intraclusterdistances of ratees that are significantly smaller than the intercluster distances. (Mean difference of intra- and interclusterdistance is significant on the p<5%-level)

Nevertheless in a few cases the inversed effect was found, were managers were rated inconsistently, which means that raters have a very different view on the overall competence level of one and the same person. This effect was found by manager C, I and K. Manager K can be considered as an outlier since the HR manager involved in the present study reported an interpersonal problem of rater 1 and manager K, which may have lead to a strong severity effect. The mean of intra- and interclassdistances in the table below had been adjusted in the data set by eliminating the major outlier of manager K.

| **managers** | **intracluster distances** | **intercluster distances** | **p-value** | **sign.** |
|---|---|---|---|---|
| mg A | 1.17 | 1.26 | 0.39 | n.s |
| mg B | 0.82 | 1.21 | 0.19 | n.s |
| *mg C* | *2.98* | *2.80* | *0.38* | *n.s* |
| mg D | 0.41 | 1.07 | 0.06 | n.s |
| mg E | 0.62 | 1.49 | 0.03 | p<5% |
| mg F | 0.71 | 1.27 | 0.06 | n.s |
| mg G | 0.82 | 1.14 | 0.19 | n.s |
| mg H | 1.10 | 1.22 | 0.39 | n.s |
| *mg I* | *2.07* | *1.60* | *0.07* | *n.s* |
| mg J | 1.33 | 1.45 | 0.35 | n.s |
| *mg K* | *4.06* | *2.47* | *0.00* | *P<0.1%* |
| mg L | 1.46 | 1.31 | 0.35 | n.s |
| mg M | 1.20 | 1.30 | 0.40 | n.s |
| mg N | 0.82 | 1.36 | 0.12 | n.s |
| mg O | 0.97 | 1.22 | 0.26 | n.s |
| *Mean* | **1.18** | **1.41** | **0.016** | **p<5%** |

Table 4: Ratee-Effect: Intra- vs. interclass Distances of city block map

As a conclusion of table 4 and table 1 we have found an interesting effect which can be attributed as a general halo effect deriving from an overall competence score. When taking the overall competence score into account using city block distance as a proximity measure, interrater reliability is higher than when we compare only the competency profiles in terms of correlation coefficients used in the NMDS map based on Pearson correlations. This finding is considered as rather important, since it shows that performance ratings are quite accurate in terms of interrater reliability on a holistic level, but cannot be taken for granted on a dimensional level.

As a final result we emphasize the confirmation of the second hypothesis of a holistic model by comparing correlation coefficients on a holistic and dimensional level.

| | r1 | | r3 | | r2a | | r2b | | r2c | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Corr Dim | Corr overall | Corr Dim | Corr overall | Corr Dim | Corr overall | Corr Dim | Corr overall | Corr Dim | Corr overall |
| **r1** | 1.00 | 1.00 | | | | | | | | |
| **r3** | 0.36 | 0.83 | 1.00 | 1.00 | | | | | | |
| **r2a** | 0.42 | 0.57 | 0.07 | 0.09 | 1.00 | 1.00 | | | | |
| **r2b** | 0.47 | 0.82 | 0.18 | 0.46 | x | x | 1.00 | 1.00 | | |
| **r2c** | 0.47 | 0.89 | 0.45 | 0.92 | x | x | x | x | 1.00 | 1.00 |

Table 5: Mean of Pearson correlation coefficients based on dimensional competency profiles and holistic overall competency scores

When correlating the 15 ratees with each pair of raters once on a dimensional and once on a holistic level (i.e. overall competency level), we have found significantly higher correlations on the overall competency level than on the dimensional correlation model. The table below illustrates how iterraterreliability on a dimensional level (Mean Corr. Dim= 0.35) is significantly smaller than the correlations on overall competence scores. (Mean Corr. overall=0.65). Using a two sided t-test for independent samples we have found this effect to be significant on the 5% level (p=0.017). The implications of these finding will be discussed in the final chapter.

## 3.5 Discussion

The present study had two main purposes. On the one hand our aim was to demonstrate the strong rater bias in performance ratings using a new methodological approach by applying multidimensional scaling which visualizes the rater dependency considering the whole variance in the data. On the other hand the study shows that interrater reliability is rather given on a holistic than a dimensional view. First the main and robust effect of the rater bias shall be discussed:

Ideally, the rating variance associated with the performance of the rate would be large relative to the variance associated with biases of the rater. In other words, what is being rated should account for more variance than does who is doing the rating. The results presented in this study shows that this is not true. Our findings parallel those of Lance (1994), who concluded that “ratings were stronger reflections of raters’ overall biases than of true performance factors (p.768), and are surprisingly consistent with a generalizability theory analysis of ratings by Greguras and Robie (1998). Greguras and Robie estimated several sources of ratings variance and then compared the total

variance associated with ratees (i.e., rater main effects and rate x rater interaction effects) with the variance associated with ratees (i.e., true score effects). For superior ratings, as we have them in our study, Greguras and Robie found that the rater effects were 1.17 times as large as ratee effects. Raters seem to agree on an overall performance score, when it comes down to the question who performs better on a specific competence, raters show less coherence. This finding is consistent with those of Viswesvaran (1993), who found, as already mentioned in the introduction of this article, support for a strong general performance factor in 25 distinct measures of job performance. Scullen, Mount and Goff (2000) mention that it is not clear why the general factor exerts a more powerful influence on ratings than do the dimensional factors. We assume that each rater has a very different perception of what is important to succeed in a job and that this estimation of importance may influence their ratings on a dimensional level.

In order to understand why rater biases are so strong, one must consider the idiosyncratic nature of performance ratings. First, raters form different perspectives by focusing their attention on different aspects of the ratee's performance. Second, as mentioned earlier on, raters from different perspectives might attend to the same aspects of performance but attach different weights to them. This explains the low dimensional correlations between the different raters. Third, raters from different perspectives often observe different samples of a ratee's behavior. Thus, ratings might differ across perspectives because of real differences in the behaviours that are observed. Borman (1997) even argues that differences across perspectives of raters may not be biases at all. Instead they may reflect portions of the true performance criterion space that are unique to ratings from each perspective.

The authors of this study introduce an alternative interpretation arguing that the raters are over challenged with the rating format of a conventional performance appraisal form where performance level and profile information on different dimensions have to be considered simultaneously when rating different ratees. That this interpretation is a valid possible explanation is also shown when comparing the NMDS maps based on Pearson Correlation and City block distances. The city block distances take into account the absolute level information and confirm the strong rater effect when combining performance level and profile information. However, in both maps the rater bias occur which leads to the recommendation of disentangling performance level and profile information. This could be reached with a forced-choice approach

rating only on a dimensional level and in a second step asking raters to estimate the overall performance level from a holistic point of view.

This proposed procedure was examined in further research by the authors of the present study.

*Practical Implications*

Taken as a whole, the findings presented in this study suggest some potential problems and pitfalls in the use of performance ratings for research and administrative purposes.

Performance ratings are used in practice to make decisions concerning pay raises, promotions, and terminations. The results illustrated on the NMDS maps show that the main differences of performance ratings are associated with biases of the rater than with the performance of the ratee. Although, this may have already been known in a general sense, previous research could not demonstrate this effect by visualizing performance data on a two dimensional map. Traditional research approaches mainly use analysis of variance or confirmatory factor analysis when dealing with multitrait-multirater data. The author of the present study argues that NMDS is a much more effective method to demonstrate rater biases than rather abstracts proportions of explained variance or correlation coefficients.

In scientific research, corrections for unreliability can account for the effects of measurement error; however, there is no analogous correction factor than can be used in organizations to eliminate the effects of idiosyncratic rater bias. The obvious implication of our finding is that decision makers should be aware of the impact of idiosyncratic bias and attempt to control its effects. This could be done by seeking a variety of types of performance information, possibly including objective measures or ratings made by multiple individuals.

The results found in this study also illustrate that organizations can gain significant benefit from using multirater systems. Generalizability theorists (Cronbach, Gleser, Nanda & Rajaranam, 1972) have shown that if ratings are averaged across n raters, each of the error components divided by n, whereas the true variance components remain unchanged. This effectively increases the proportion of true variance. Of course, the larger the error components are, the greater is the advantage of using multiple raters. Because the error components, especially idiosyncratic variance, are

large, averaging across several raters can significantly reduce the effects of bias and random error.

Further research has to emphasize on the causes or the nature of individual (i.e., idiosyncratic) and perspective related effect, and research of this type is clearly needed.

The second main robust finding concerns the low dimensional reliability of performance data. The findings reported here do not support a dimension oriented practice in multisource feedback systems. Specifically, the ratings that managers receive from different raters on each dimension often are compared directly in these systems (Dunnette, 1993, London & Smither, 1995; Tornow, 1993). The results reported are a hint that comparisons such as these are not legitimate. This perspective becomes important when job performance is used as a source of developmental feedback. The results support the conclusion that the competency based questionnaire shows important variance across the 5 raters, which means we cannot assume that the same underlying competency constructs were being measured by the raters. Differences of the ratings can therefore not be attributed to the true differences in competency levels, but have more to be considered as an artefact of a rater bias.

On a practical level this means, that when applying multirater performance instruments, we may assume reliability concerning overall competence score but we should be careful when making conclusions on developing certain competencies since reliability of competency dimensions can not be taken for granted.

*Limitations*

The generalizability of our results is bounded by our focus on one organization and a non validated competency model. Although the competency modelling was based on the general approach by the proponents of competency modelling (i.e., Shipmann, 2000), future studies are needed based on a validated competency model.

Another question mark is put on the capability of certain raters to make concise ratings due to lack of exposure to the ratees. One of the raters is head of human resources and therefore not in direct day to day interaction. This might also have influenced the validity of the data.

A final limitation is the relatively small sample size of our study. However, it is important to remember that samples like the ones we employed involve upper

management level whose time is precious to the organization. In fact, it can be argued that our sample size is representative of the typical subject matter experts.

*Conclusion*

To sum it up in a nutshell, the main contribution of this study was to enhance our understanding of competency based performance ratings visualizing strong rater effects using Nonmetric Multidimensional Scaling (NMDS). The results quantify rater and ratee caused differences in competency level by comparing mean intra- and interclusterdistances of raters and ratees. Applying a form of a cross- and nested design we have found that in general idiosyncratic rater effects are significantly more responsible for explaining differences in performance data than the actual differences in job performance caused by the ratee. We have also found that when considering the overall competency score the interrater reliability is higher than when only comparing the competency profiles between different managers. This implies that competency based ratings might be applied on a holistic view for selection purposes but should be considered with precaution when deviating developmental feedback and -programs based on single competency dimensions.

## 3.6 References

Borman, W. C. (1997). 360° ratings: An analysis of assumptions and a research agenda for evaluating their validity. *Human Resource Management Review*, 7, 299-315.

Campbell, D.J., & Lee, C. (1988). Self-appraisal in performance evaluation: Development vs. evaluation. *Academy of Management Review*, 13, 302-314.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Cardy, R. L., & Dobbins, G.H. (1994). Performance appraisal: Alternatives perspectives. Cincinatti, OH: South-Western Publishing Co.

Cascio, W.F. (1991). Applied psychology in personnel management (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Cleveland, J.N., Murphy, K.R., & Williams, R.E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Jounral of Applied Psychology*, 74, 130-135.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, 90, 218-24.

Conway, J.M. , & Huffcutt, A.I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and selfratings. *Human Performance*, 10, 331-360.

Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Daub, S. (2001). Similartiy Simulation – wie man den code des Globalurteils knackt; Dissertation, Zürich, 2001.

Dunnette, M.D. (1993). My hammer is your hammer. *Human Resource Management*, 32, 373-384.

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology*, 83, 960-968.

Harris, M.M, & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, 41, 43-62.

Läge, D (2001). Ähnlichkeitsbasierte Diagnostik von Sachwissen. Habilitationsschrift, Universität Zürich, Zürich.

Lance, C. E., & Woehr, D.J (1989). The validity of performance judgments: Normative accuracy model versus ecological perspectives. In D.F. Ray (Ed.), *Proceedings of the Southern Management Association* (pp.115-117). Starkville, MS: Sothern Management Association.

Lance, C.E., Teachout, M.S. & Donnelly, T.M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology*, 77, 437-452.

Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology*, 79, 332-340.

Landy, F.J., & Farr, J.L. (1980), Performance rating. *Psychological Bulletin*, 87, 72-107.

Lievens, F., Sanches, J.I., & De Corte, W. (2004). Easing the inferential leap in competency modeling: The effects of task-related information and subject matter expertise. *Personnel Psychology*, 57, 881-904.

London, M., & Smither, J.W. (1995). Can muldti-source feedback change perception of goal accomplishment, self-evaluations, and performance rated outcomes? Theory based applications and directions for research. *Personnel Psychology*, 48, 803-839.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitraitmultimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112, 165-172.

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology*, 65, 507-516.

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology*, 65, 507-516.

McClelland, D. C. (1973). Testing for competence rather than for "intelligence". *American Psychologist*, 28, 1-14.

Motowidlo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and conceptual performance. *Human Performance*, 10, 71-83.

Mount, M.K. (1984). Psychometric properties of subordinate ratings of managerial performance. *Personnel Psychology*, 37, 687-702.

Murphy, K.R., & Cleveland, J.N. (1995). Understanding performance appraisal: Social, organizational, and goal based perspectives. Thousand Oaks, CA: Sage.

Raven, J., & Stephenson, J. (Eds.). (2001). Competence in the Learning Society. NewYork: Peter Lang.

Shippmann, J. S., Ash, R. A., Battista, M., Carr, L., Eyde, L. D., Hesketh, B., Kehoe, J., Pearlman, K., & Sanchez, J. I. (2000). The practice of competency modeling, *Personnel Psychology, 53*, 703-740.

Scullen, S.E., Mount, M.K., Goff, M., (2000). Understanding the latent structure of job performance ratings, *Journal of Applied Psychology*, 85, 956-970.

Scullen, S.E. (1999). Using confirmatory factor analysis of correlated uniquenesses to estimate method variance in multitrait-multimethod matrices. *Organizational Research Methods*, 2, 275-292.

Schmitt F.L. & Hunter J.E., (1998). The validity and utility of selection methods in personal psychology: Practical and theoretical implications of 85 years of research findings. *Pychological Bulletin*, 124, 262-274.

Shippmann, J.S., Ash R.A., Battista, M, Carr, L, Eyde, L.D, Hesketh, B. et al. (2000). The practice of competence modeling. *Personnel Psychology*, 53, 703-740.

Spencer, L. M., & Spencer, S. M. (1993). Competence at Work. New York: Wiley.

Tornow, W.W. (1993). Editor's note: Introduction to the special issue on 360-degree-feedback. *Human Resource Management*, 32, 211-219.

Viswesvaran, C. (1993). Modeling job performance: Is there a general factor? Unpublished doctoral dissertation,University of Iowa, Iowa City http://handle.dtic.mil/100.2/ADA294282.

Wherry, R. J., Sr., & Bartlett, C. J. (1982). The control of bias in ratings:a theory of rating. *Personnel Psychology*, 35, 521-551.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, 9, 1-26.

# 4 Comparison of forced-choice versus multipoint Likert scales in performance appraisal

## 4.1 Abstract

The purpose of this study is to show how a forced-choice format in the questionnaire is a valuable alternative to standard job performance ratings eliminating methodologically the halo effect at the construct level. A competency based questionnaire on a multipoint Likert response format is compared to a forced-choice method based on ordinal data. Nonmetric Multidimensional Scaling (NMDS) was applied to show rater and ratee-caused effects in performance appraisal to demonstrate the advantage of using a forced-choice format over multipoint Likert-type items. The implication of separating overall job performance and dimensional job performance in order to increase quality of personnel evaluation is critically assessed.

## 4.2 Introduction

Although it appears that many assume performance appraisal is quite accurate, there is considerable evidence, that the types of subjective judgements often involved in performance appraisals may be subject to systematic sources of inaccuracy. Because of its important implications in personnel selection and in the development of theories of work and work performance, halo error in ratings of job performance has been studied for nearly 100 years. Viswesvaran and Schmidt (2005) approved in their meta-analysis the theory of a general factor which explains 60% of total variance in job performance ratings.

Given that formal performance appraisal systems serve a variety of important functions within organizations (Cleveland, Murphy, Williams, 1989) and that most organizations ratings are obtained only from employees` supervisors (Bernardin & Villanova, 1986), it is not surprising that a great deal of research has focused on the psychometric quality of supervisory ratings. Unfortunately, a number of studies conducted over the past decades indicate that supervisory ratings often are plagued by a host of potential problems, including halo, leniency, intentional manipulation, and race, gender, or age bias biases (see Cardy & Dobbins, 1994, Cascio, 1991; Landy & Farr, 1980, for reviews). Especially halo error has been studied thoroughly for over 100 years. Viswesvaran and Schmidt (2005) have used a database integrating 90 years of empirical studies reporting inter-correlations among rated job performance dimensions to test the hypothesis of a general factor in job performance. After controlling for halo error and

three other sources of measurement error, there remained a general factor in job performance ratings at the construct level accounting for 60% of total variance.

Recognizing these limitations in job performance ratings, researchers turned their attention to examining alternative rating sources such as peer, subordinate, and self ratings.

Perhaps one of the most consistent findings in the empirical literature on performance appraisal systems is that the ratings obtained from different sources generally do not converge. The inter-correlations among the ratings provided by different type of raters tend to be moderate at the best. For example, in a meta-analysis Harris and Schaubroeck (1988) reported mean self-supervisor, self-peer, and peer-supervisor rating correlations (corrected for unreliability) of .35, .36, and .62, respectively. Similarly, Mount (1984) reported mean supervisor-subordinate and subordinate-self rating correlation of .24 and .19, respectively. Finally, Conway and Huffcutt (1997) reported that the correlations (Corrected for unreliability) among ratings made by self, peer, supervisor, and subordinate raters ranged from a high of .79 (supervisor –peer) to a low of .14 (subordinate-self).

The average ratings provided by different rating sources also tend to differ. Thornton (1980) reported that a „preponderance of studies show that individuals rate themselves higher than they are rated by comparison groups“(p.265). Mount (1984) found that self ratings tend to be higher than supervisor ratings, which in turn tend to be higher than subordinate ratings. Harris and Schaubröck (1988) found that self ratings averaged .70 SD higher than supervisor ratings and .28 SD higher than peer ratings. Peer ratings averaged .23 SD higher than supervisor ratings. Although consistent with Mount´s findings, none of these differences were significant, however. Harris and Schaubbröck attributed this to the large variance in effects across studies included in their meta-analysis.

A number of different explanations have emerged to account for why different types of raters do not agree in their ratings. Campbell and Lee (1988) argued that different rater groups may have different conceptualizations of what constitutes effective performance in a particular job. Murphy and Cleveland (1995) suggested that raters differ in their opportunity to observe any given individuals´ work behaviour and that these differences in perspective may account for disagreements among their ratings. Similarly, Lance, Teachout, and Donnelly (1992) noted that modest correspondence among ratings from different types of raters may be due to the fact that they are exposed to only moderately overlapping sets of ratee behaviour. This „ecological perspective“ on rating source differences (Lance & Woehr, 1989) suggests that strong correspondence among ratings from different sources should not be expected. Finally, Campell and Lee (1988) as well as Cardy and Dobbins (1994) described a number of

processes that may lead to disagreements among the ratings provided by different types of raters. These processes include well-established attributional tendencies, such as the self-serving attributional bias and the actor-observer effect, as well as motivational and informational differences between rating sources such as self raters needs for self enhancement and differences in social comparison information available to self raters and their supervision (Fahr & Dobbins, 1989).

After all Mount, M.K., Judge, T.A., Scullen, S.E., Sytsma, M.R., & Hezlett, S.A. (1998) suggested, that the rater level is not a decisive factor when explaining variance in performance data. Their results show that method variance in multitrait-multirater (MTMR) data is more strongly associated with individual raters than with the rater's level.

Although the well confirmed finding of a general factor which leads to biased performance data as well as the compelling explanations for why performance appraisal ratings from alternative sources differ, have lead to a better understanding of the neuralgic points of performance data, another fundamental question is raised in this present study, questioning the influence of the response format of a rating instrument when evaluating job performance data. Based on the stated findings of strong rater effects, we argue in this study that these strong rater biases are mainly caused by the questioning format. Most multisource feedback instruments in industry use multipoint Likert scales which can not avoid measuring an idiosyncratic level of overall job performance which can also be considered as a general factor influencing dimensional ratings. This kind of multipoint Likert response format asks the rater to complete two tasks simultaneously and may lead to a cognitive bottle neck. One task is to keep in mind overall job performance over different ratees while a simultaneous second task is to draw a profile of different competencies within an individual. The advantage of a forced- choice format is that the first task of keeping in mind the overall performance score over several individuals drops out, which makes the task for the rater a lot easier.

One purpose of this study is to use new methods to illustrate that there is a general factor among different raters and to show how this bias of a general factor can be reduced by using a forced-choice response format. Our hypothesis is that the rater-effect preponderates in a Likert response format compared to the forced-choice format. Our operationalized hypothesis is that the congruence among ratees on a dimensional level is significantly higher using a forced-choice format rather than using multipoint Likert response format.

## 4.3 Method

*Participants and design*

The study included a sample of 15 upper level managers (ratees) of a large, Swiss utility company in Switzerland. The managers represented a variety of functional areas, such as power generation, power delivery, and customer operations. Because of concerns for anonymity, no information regarding raters' demographic characteristics was collected. The 15 managers were rated in a multisource rating design of 3 different functional roles: Chief Executive Officer (rater 1), 3 members of the executive board as direct line superiors of the departments “energy”, “grid” and “finance” (raters 2a, 2b, 2c) and Head Human Resources (rater 3). This is a typical setting which we encounter in practice. Rater 1 and 3 rated all ratees while raters 2a, 2b and 2c rated only the ratees in their department. This corresponds to a mixed design of cross and nested ratings. In cross rating systems, each rater rates the performance of all ratees. In nested ratings systems, each ratee’s performance is rated by different raters. This distinction is important because the amount of idiosyncratic ratings variance is expected to be larger in a nested design than in a crossed design. The reason is that each rater in a nested design may exhibit a different degree of leniency. Leniency errors refers to a rater’s tendency to assign ratings that are generally higher (or lower) than are warranted by the ratee’s actual performance. Thus, rater leniency differences introduce an element of variability into nested designs that is not present in crossed designs. To the knowledge of the author of the present study, the magnitude of the difference in idiosyncratic variance between these two different types of designs is not yet known, largely because research has failed to acknowledge the distinctions between the two types of designs. In this study, both halo and leniency differences contribute to idiosyncratic rating variance. Because the intention in this study was to quantify the overall effect of rater bias, the distinction of leniency and halo effect was not attempted.

*Procedure*

The data for this study were collected during the administration of a competency based evaluation of the first management level. The competency based questionnaire was identical for each rater group, and it contained 51 behaviourally oriented items that were designed to measure 17 dimensions of managerial competencies. According to Lievens, Sanchez & De Corte (2004) competency modelling ties the derivation of job specifications to the organization’s strategy, which, together with nonstrategic job requirements, are used to generate a “common language” in the form of a set of human attributes or individual

competencies. The competence dimensions included entrepreneurship, customer orientation, goal and result orientation, innovation, planning & organizing, analysis and problem solving, implementation, initiative, reasoning and decision making, resilience, leadership, integrity, influence and control, conflict management, communication, teamwork and networking skills as well as occupational and professional knowledge. These competency dimensions were the result of a benchmark study that compared researched based and practical proved[4] competency models identified by the organization as being critical for successful leadership within the company. Managers who were being evaluated as part of the multisource competency evaluation completed self ratings of their performance. After rating forms were processed, managers received feedback about their ratings in the form of a feedback report. These reports summarized the ratings that managers received on each performance dimension as well as on each item by rating source.

In addition to the competency based questionnaire format, the raters were asked to complete a forced-choice tool in a multidimensional forced-choice format (MFC). An MFC comprises 17 competency statements, each reflecting a different, independent competency dimension according the competency model presented earlier. In a computer based application respondents were asked to follow a two-step procedure.

In a first step, raters were requested to choose 8 out 17 competencies that are considered as relative strengths of the ratee.

[4] The benchmark study was based on competency research of Boyatzis (1982), Spencer & Spencer (2008), Thornton & Byham (1982), Shipman et al. (1999), the Swiss Military Leadership Institute, the Psychology Department of Applied Cognitive Sciences at University of Zürich (Zuber, Matthys & Läge, 2005) and practical application of competency models used in 4 different companies in the energy branch.


Figure 1: Forced-choice procedure in performance appraisal, step 1.

In a second step raters were asked to put the 8 chosen competencies in a rank order on a one dimensional scale. The gradual scale was indicated with two anchors at each pole („competency level highly distinct“ and „competency level less highly distinct“).

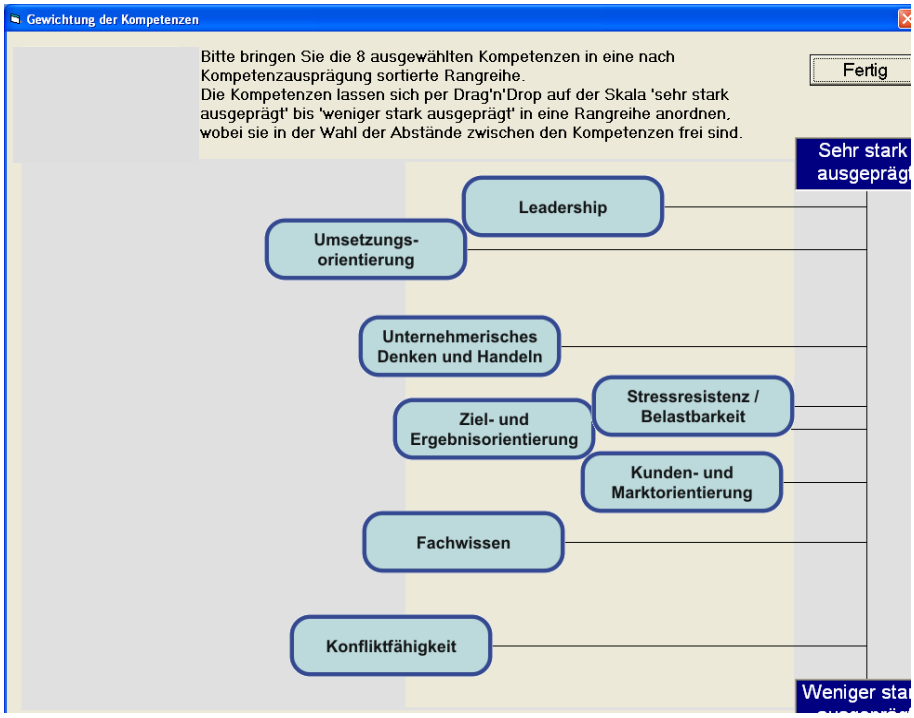



Figure 2: Forced-choice procedure in performance appraisal, step 2.

*Measures*

A brief description of each dimension on the original competency questionnaire appears in the appendix. When completing the questionnaire, raters were encouraged to think about how often the manager they were rating demonstrated the behaviour described by each item. They were asked to base their ratings on behaviours that they had observed, not on how they believed a manager might behave. Ratings were made on an 8-point Likert-type scale that was segmented in 4 behavioural anchors from "poor" to "excellent".

Forced-choice response formats produce ordinal data. According to the forced-choice procedure described earlier in this study, each competency construct of each ratee obtains a value between 1 and 9. Applying this forced-choice procedure the data collected is to be considered as ipsative measures. The defining characteristic of ipsative measurement is that the total score on the instrument (sum of scores across the scales) is a constant for all examinees. Ipsative measures require the respondent to distribute this fixed number of points across the constructs assessed by the measure, such that scale scores are distributed about the individual's mean over all the constructs. Ipsative measures, therefore, provide information on the rank ordering of the constructs for the particular individual being assessed (i.e., intraindividual-differences assessment). In other words, ipsative scores derive meaning by comparing an individual's scores on one scale with his or her scores on the other scales included in the measure. Since all ratees have been measured with the same tool on a ranking scale from 1 to 9, the mean of all individuals is the same and we can therefore compare interindividual differences on a construct level by using correlation measures.

*Analysis*

As explained in one of our previous studies[5], we used Nonmetric Multidimensional Scaling Technique (NMDS) rather than the commonly used confirmatory factor analysis in order to illustrate the latent structure of job performance ratings. The data collected with the two different instruments based on competency scores on each of the 17 dimensions were copied into correlation matrices and then analyzed by robust NMDS (by means of the ROBUSCAL algorithm (Läge, 2001). NMDS is a nonmetric algorithm, which interprets the competency scores on an ordinal level. In an iterative process, the configuration, which best corresponds to the proximities between the rated managers is approximated. The target configuration describes an n-dimensional space, which in this study resulted in two dimensions (the cognitive map). The cognitive map depicts the cognitive relational structure of a subject in the

[5] Unpubl. Ph.D. thesis, University of Zürich, Institute of Psychology, Dept. of Applied Cognitive Sciences, 2010.

form of a Euclidian space, where a small distance between the objects (i.e. competency profile of a manager) corresponds to a high competence profile similarity and vice versa. One may correctly argue that in correlation measures the absolute competency level does not have any influence on the similarity between two objects. We therefore computed also cognitive maps taking into account the absolute score (i.e. overall competency score) using city block distances, $d_{ik} = \sum_{j=1}^{p} \left| x_{ij} - x_{kj} \right|$, in order to calculate the matrices of competency profiles. As in one of our previous studies (see chapter 3) we used multiple regression (property fitting) to fit in a regression line of the overall performance level into the two-dimensional NDMS map. This method was used to illustrate a strong rater bias when applying multipoint Likert response format. A high multiple correlation of the dimension in the NMDS configuration indicates a high level of explanation of total variance.

In order to compare the two measurement instruments regarding the congruence on a rater- and ratee level, the distances were extracted from the Euclidian maps. Intracluster distances within rater and ratee were then calculated and analyzed for significant differences using common inferential statistics.

## 4.4 Results

One of our main results of this study is represented in a NMDS map (see figure 3), that has been calculated based on city block distances which takes the absolute level of performance appraisal into consideration. This map depicts the similarity of ratee´s competency profiles across all different raters. Before looking at the map with regard to content, it is necessary to explain an internal quality measure of the ROBUSCAL algorithm: the standardized stress value (Läge, 2001) is a measure of how well the algorithm was able to translate the similarity judgments into an n-dimensional map. Thus, it is also an indicator of the level of consistency of the competency profiles across the raters and consequently provides a marker of the interpretability of the cognitive map. According to literature (Borg & Groenen, 1997 and Purkhardt & Stockdale, 1993) the stress value of this mean map of 0.118 are more than acceptable, and thus interpretable.

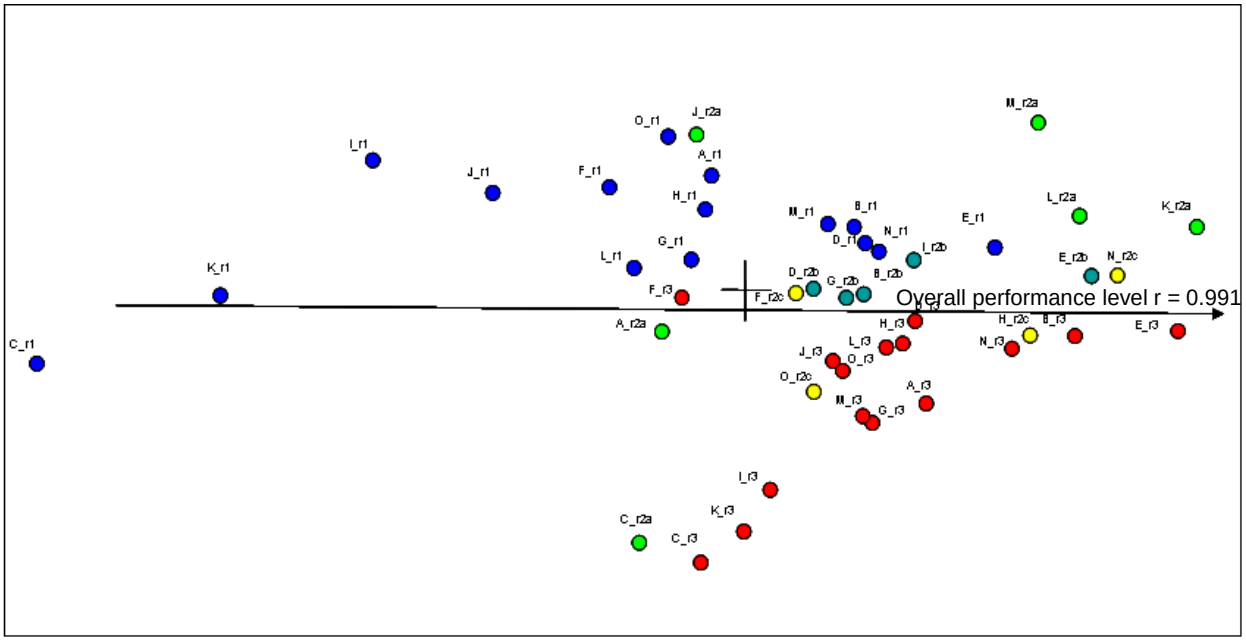



Figure 3: NMDS-map using city block distances

The alphabetical marked dots represent competency profiles of rated managers based on the score of 17 competency dimensions. Each rater is marked with a different colour which helps to illustrate the strong rater effect in the data. The similarity of the profiles is mainly due to the difference in overall performance level. Whereas rater (r1) has generally given low performance scores (leniency effect), rater (2a) was rather "gentle" in his overall ratings.

Overall the mean and standard deviation of the intracluster distances among raters is smaller (Mean = 1.29, SD = 1.15) than the intracluster distances among the ratees (Mean = 1.36, SD = 1.17). Even though these differences are slightly not significant, there is some evidence that on a dimensional level raters are not very consistent in their ratings. This hypothesis can not be shown consequently using city block measures since the overall performance level "overrules" the effect on a dimensional level.

In order to show that the rater bias is not only due to the overall performance level, but dependent on the questioning format, we compared the intracluster distances of the two Euclidian maps, Likert response- and forced-choice format, which were both calculated on correlation matrices, rather than city block measures.

Figures 4 and 5 show the results of this comparison illustrating the rater-effect depending on the response format of the measuring instrument.

**Rater-effect depending on the response format**

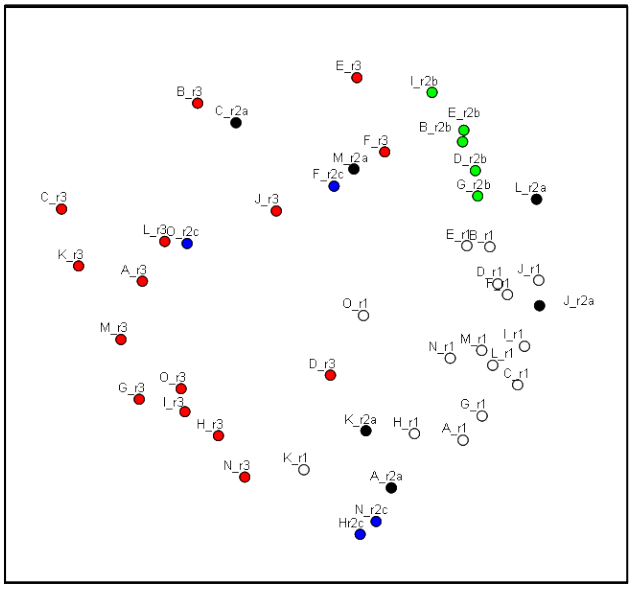



Figure 4: Multipoint Likert response format

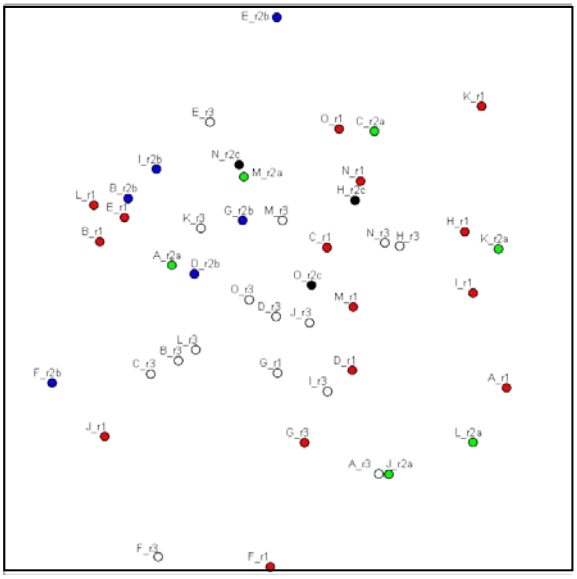



Figure 5: Forced-choice response format

These results posit that covariation in ratings using multipoint Likert questionnaires is mainly associated with individual raters and scarcely associated with the real traits (competency scores) of the ratees. The relative position of a ratee in a two dimensional Euclidian map based on correlation measure is highly determined by the rater's view. Each rater has an overall impression in which competencies the ratees have their strengths and weaknesses. This can be shown in the figure 6 illustrating the rater's mean scores on each competency construct across all ratees.

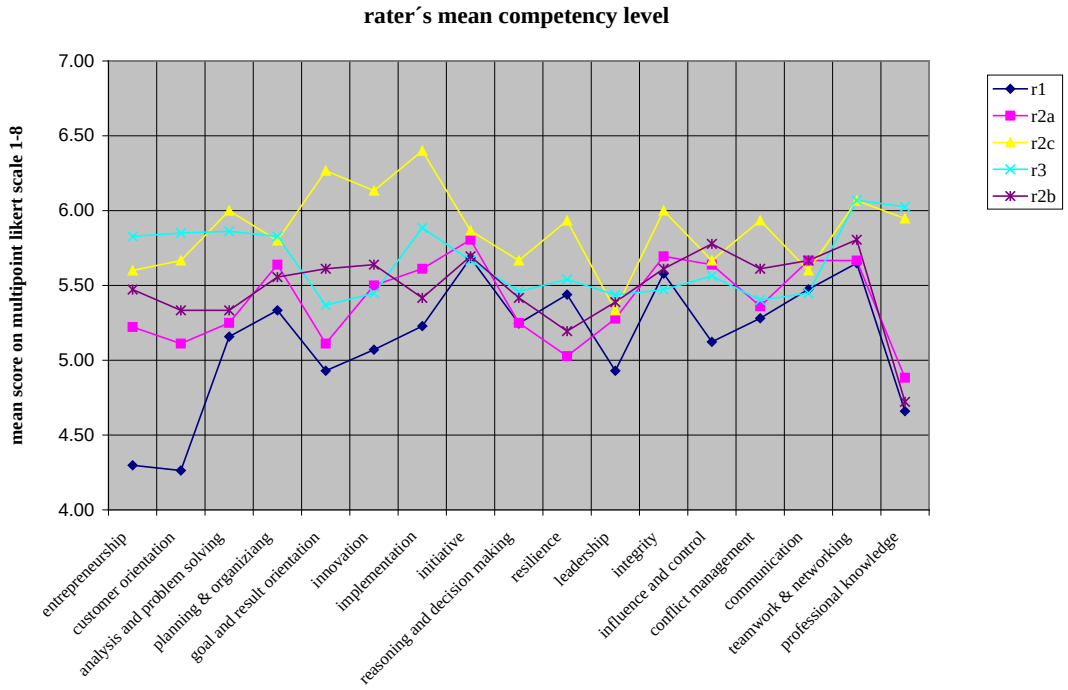


Figure 6: Rater's mean on 15 competencies across all ratees

This overall view on the strength's and weaknesses in the mind of the different raters lead to high congruence within one rater, but not to high congruence within the competence profile of one ratee rated by different raters.

This rather strong rater bias observed in the Euclidian map of figure 4 does not apply in forced-choice response format shown in figure 5. Here the similarity of competency profiles does not depend as much of the rater. The congruence of profiles within a rater is not given.

In order to investigate weather the congruence at construct level among ratee's is higher than the congruence level within the different raters, we marked the competency profiles represented as dots in the two-dimensional maps according to the different ratees for both response formats (i.e. figure 7 and 8.).

**Ratee effect depending on the response format**

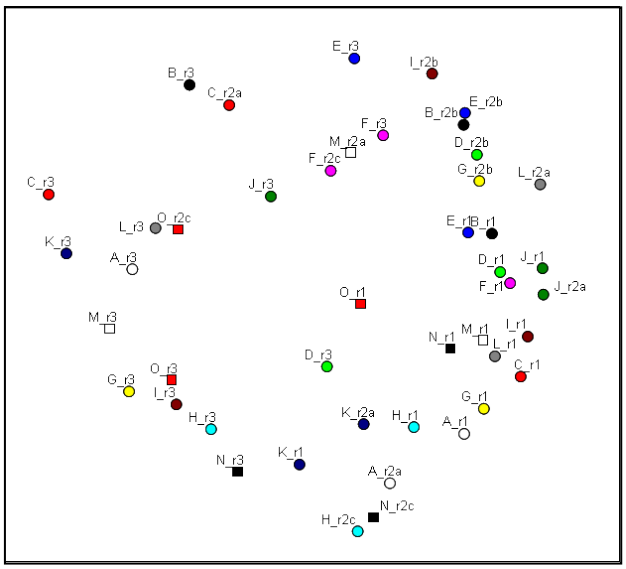



Figure 7: Likert-response format

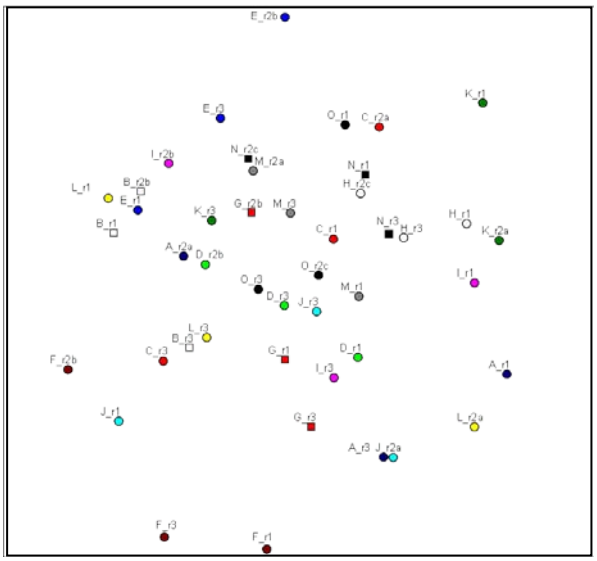


Figure 8: Forced-choice response format

At a first glance when comparing the two NMDS maps, the congruence among ratees does not differ prominently in the two different response formats. We expect a switch of the effect when comparing intracluster distances with the two different rating formats. Due to the strong rater effect caused by the measurement of absolute scores in Likert response formats, we expect smaller intraclass distances for raters than for rates. When we use forced-choice method, where the absolute score does not have any significance, we expect a switch of this effect, finding smaller intracluster distances within ratees than within raters. When putting the intracluster distances among ratees and raters into numbers, we find an effect confirming our

hypothesis of higher congruence among ratees than among raters when using a forced-choice response format.

| Likert Response Format | | | | | Forced Choice Format | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Intracluster rater | Intracluster ratee | p-value | significance | | Intracluster rater | Intracluster ratee | p-value | significance |
| Mean all rater | 0.91 | 1.32 | 0.005 | p < 0.1% | Mean all rater | 1.35 | 1.10 | 0.041 | p < 5% |
| r1 | 0.66 | 1.32 | 0.00 | p < 0.01% | r1 | 1.47 | 1.10 | 0.0034 | p < 0.01% |
| r2a | 1.41 | 1.32 | 0.31 | n.s | r2a | 1.51 | 1.10 | 0.03 | p < 5% |
| r2b | 0.30 | 1.32 | 2.6 E-06 | p < 0.01% | r2b | 1.04 | 1.10 | 0.32 | n.s |
| r2c | 1.45 | 1.32 | 0.65 | n.s | r2c | 1.22 | 1.10 | 0.39 | n.s |
| r3 | 1.12 | 1.32 | 0.02 | p < 5% | r3 | 1.20 | 1.10 | 0.29 | n.s |

Table 1: Intracluster- and intercluster distances of Likert-response and forced-choice format

As mentioned above the rater effect is stronger using Likert response format. When we compare the intraclusters among the raters in the two different measurement formats using t-test for dependent samples, we find highly significant (p<0.01%) smaller intracluster distances among raters (mean = 0.91) using Likert response format than intracluster distances among raters using forced-choice format (mean = 1.35).

When we look at the intracluster distances among ratees, the effect switches into the other direction. Using t-test for the mean intracluster differences among ratees in the two different measurement settings, we obtain smaller intracluster-distances with forced-choice data (mean = 1.10) than with Likert response format (mean = 1.32). This mean differences are significant on the p < 5% level.

## 4.5 Discussion

About 90 years ago Thorndike (1920) observed that when supervisors rated their subordinates, the correlations among performance dimensions were "higher than reality" (p. 25) and "too high and too even" (p. 27). Research conducted since that time has further documented the ubiquitous phenomenon of method effects in performance ratings, and has shown that such effects represent one of the largest sources of error in performance ratings (Cooper, 1981). However, an unanswered question in this body of research is whether such effects are associated with the questioning format of the performance ratings.

We tested two types of questioning formats that hypothesized different congruence among raters and ratees on a construct level. The finding of central interest is that method variance in performance ratings is associated with strong rater bias when using multipoint Likert questionnaires rather than forced-choice format.

Research reviewed earlier indicated that method effects in multitrait-multirater data are commonplace analysis. This is consistent with Becker and Cote's (1994) recommendation, "…if boundary conditions are not a problem, confirmatory factor analysis is generally preferable to alternative methods" (p. 635). Nevertheless, we suggested earlier Nonmetric Multidimensional Scaling Technique as an alternative to factor analysis in order to illustrate rater effect on a holistic point of view. We could demonstrate that rater effect is strongly associated with their view of overall performance among the different ratees which highly influence their ratings on a dimensional level. We could also show that raters have an idiosyncratic model of strength and weaknesses in their mind that they apply over all ratees. Rater 1 for instance has systematically rated competencies as integrity, initiative as well as team- and networking skills as strengths whereas entrepreneurship, goal and result orientation as well as leadership have been rated as overall weaknesses of his subordinates. These holistic views of management, as we call them at this point, are most likely responsible for a strong rater effect shown in the NMDS maps. We argue that reason for this effect relies on a cognitive bottleneck when having to deal with two tasks simultaneously. First rater have to keep in mind their overall ranking (i.e. sum scores over all competencies), second they have to make a profile of strengths and weaknesses in terms of managerial competencies while keeping the ranking of their overall performance score upright. Based on our findings, we suggest that these two tasks need to be separated in order to reduce complexity and cognitive overload. We recommend to rate performance data in two steps. The first step is to make an overall ranking of performance score on a holistic level comparing ratees directly with each other. This could be practically implemented by using our earlier presented forced-choice tool, placing ratee’s according to their overall performance level instead of competency dimensions into a ranking order on a continuous two-poled ideal scale from low to high performance. In a second step raters are asked to focus only on the competency profile of a ratee placing competencies according to the strengths and weaknesses on a two-poled ideal scale. Following these two simple steps, holistic performance level and profile information on construct level can be assessed without reducing validity through strong rater’s halo effect.. According to respondents the task of direct comparison of the ratees is easier than the simultaneous integration of overall performance score and dimensional profiling of common performance appraisal forms.

Coming back on this present study, one may argue the use of a random dimensional taxonomy of our competency scales constrained our ability to derive higher congruence among ratees on a dimensional level. In order to address this issue, we conducted a principal components

analysis (with varimax rotation) of the 17 competency dimensions. It revealed a 3-factor structure which corresponded very closely to Wunderer´s (2007) three factors. The three factors and the corresponding competency dimensions and factor loadings were as follows: Factor 1-Human Relations: Leadership (.67), Integrity (.78), Conflict Management (.823), Communication (.71), influence and control (.55); Factor 2- Implementation: Problem solving (.833), planning and organizing (.74), goal- and result orientation (.81), resilience (.61), occupational knowledge; Factor 3-Creation competence: entrepreneuring (.71), Innovation (.46), customer orientation (.86), networking (.65). Several of the dimensions had cross loadings greater than .40 on other competency dimensions: Planning and organizing (.40), Innovation (.51), resilience (.53) and networking (.54) loaded on Factor 1-Human Relations. Entrepreneurship (.47), Innovation (.44) and reasoning & decision making (.57) and influence & control (.46) loaded on the implementation factor. The percentage variance accounted for by three factors was 71.90. These results indicate that Wunderer's (2007) taxonomy provides an acceptable framework for examining trait affects associated with the ratings on the multidimensional competency scales. An important implication of these findings is that the lack of rater congruence is not due to the inappropriateness of the competency framework used in this study. One other implication of our findings is that researchers and practitioners may find Wunderer's model of managerial performance to be a useful taxonomy for investigating managers' performance.

The finding that the method used collecting performance data has influence on the congruence at construct level among ratees, raises several issues for future research. Although we know that there is unique variance associated with each rater's ratings due to their different measurement level, we cannot necessarily conclude that each is accounting for unique true score variance. Thus, we cannot draw conclusions about the relative accuracy of ratings provided by different raters. Another issue is whether the present results would generalize to ratees in occupations other than management (where most of the raters are themselves managers). Here further research is needed addressing different target groups than managers.

The present results have several practical implications also in terms of use of multiperspective performance appraisal systems, such as 360°-Feeddback. At the most basic level the results provide support for the implicit premise underlying most 360-degree systems. That is, because ratings from each rater, regardless of level, appear to capture unique rating variance, it is important to include multiple raters in the process rather than relying on the results of a single rater, such as the boss. Our results show that each rater's ratings are different enough from those of other raters to constitute a separate method. The implication of this for

multiperspective feedback reports is that information should be displayed separately for each individual rater. Displaying information in this way allows the ratee to examine the pattern of each rater's ratings across the skills to determine relative strengths and weaknesses. This can of course also be done in an anonymous manner. For example, a finding that each of the five raters assigned their lowest ratings on the leadership competency would provide important feedback to a ratee that this skill area is a relative weakness and should be the focus of developmental planning. Because most multiperspective feedback programs are used primarily for developmental purposes where the focus is on the identification of strengths and weaknesses, information about the pattern of individual raters' ratings across traits is of critical importance. Thus, for this important use of multiperspective feedback, our results indicate that it is best to consider each rater's ratings separately.

There is one additional caveat we would like to point out regarding aggregation practices in 360-feedback reports. Even in those cases where it is appropriate to compute means within the level of raters (as for superiors), it is important to provide a measure of the dispersion of the ratings. This might include reporting the actual ratings made by the raters on the scale in order to illustrate how different individual ratings might be, or it might include the standard deviation of raters', ratings. Clearly, inferences drawn about the meaning of an average rating of 3.0 with a standard deviation of 2.0 (on an 8-point scale) are quite different from those drawn from the same average with a standard deviation of zero. Therefore, we strongly recommend that in those limited situations where it is appropriate to average ratings, information regarding the dispersion of the ratings should also be provided to facilitate interpretation of the feedback. When interpreting the results of this research, several issues should be kept in mind. First, ratings in this study were made for developmental purposes only, which raises a question about the generalization of our results to ratings provided in the context of promotion decisions. However, Kraiger and Ford (1985) investigated a similar question in their meta-analytic study, and found that rater-ratee effects were not moderated by purpose of the ratings. In light of their findings the generalization concerns expressed above may be minor. Another issue pertains to the conditions under which ratings were obtained in the study and their potential effects on the subsequent ratings. Managers' participation in the study was voluntary, indicating that they were actively seeking feedback. As in most multirater feedback systems, managers in this study selected the peers and subordinates who rated them. We do not know how these raters perceived the feedback-seeking behavior of target managers in this study, nor do we know how or if the performance of managers who

participate voluntarily differs from the performance raters who do not. Both of these issues appear to be worthwhile topicsfor future research.

Despite these potential limitations, we believe there are several features of this study that enhance its contribution to the literature. The study used a large sample of ratees who held jobs from the same job family (management) and were rated using a common instrument and for a common purpose, thereby eliminating potential confounding for any of these reasons. Further, managers were rated by seven raters from four levels, two bosses, two peers, two subordinates, and self-ratings for each ratee. This unique sample along with the use of CFA, allowed us to uncouple method effects in performance ratings that are due to individual raters from those that due to the rating level. The major contribution of the paper is the finding that method effects in MTMR data are associated more strongly with individual raters than with the rater's level.

## 4.6 References

Becker and Cote, 1994. T.E. Becker and J.A. Cote , Additive and multiplicative method effects in applied psychological research: An empirical assessment of three models. *Journal of Management*, 20, 625–641.

Bernardin, H. J., & Villanova, P. (1986). Performance appraisal. In E. Locke (Ed.), *Generalizing from laboratory to field settings* (pp. 189–206). Lexington, MA: Lexington Books.

Borg, I. & Groenen, P. (1997). Modern Multidimensional Scaling: Theories and Applications, New York: Springer.

Boyatzis, R.E. (1982). The competent manager. In A. Stewart (ed.), *Motivation and Society. San Francisco* CA.: Jossey Bass.

Campbell, D.J., & Lee, C. (1988), Self-Appraisal in Performance Evaluation: Development versus Evaluation , *The Academy of Management Review*, 13, 302-314.

Cardy, R.L., Dobbins, G.H. (1994), Performance Appraisal: Alternative Perspectives, South Western, Cincinnati, OH.

Cascio, W. F. (1991). Applied psychology in personnel management (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multi-source performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance*, 10, 331-360.

Cooper W.H. (1981). Ubiquitous halo. *Psychological Bulletin*, 90, 218-244.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology*, 74, 130–135.

Fahr, J. L. and Dobbins, G. H. (1989). 'Effects of self-esteem on leniency bias in self-reports of performance: A structural equation model analysis', *Personnel Psychology*, 42, 835-850.

Harris, M. M. & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, 41, 43-62.

Kraiger, K., & Ford, J. 1985. A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology*, 70: 56-65.

Läge, D. (2001). Ähnlichkeitsbasierte Diagnostik von Sachwissen; prof diss, University of Zurich.

Lance, C.E, Teachout, M.S. and T.M. Donnelly, T.M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology*, 77, 437–452.

Lance, C. E. & Woehr, D. J. (1989). The validity of performance judgments: Normative accuracy model versus ecological perspectives. In D.F. Ray (Ed.), *Southern*

*Management Association Proceedings* (pp. 115-117). Mississippi State, MS: Southern Management Association.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87, 72–107.

Lievens, F., Sanchez, J.I., & De Corte, W. (2004). Easing the inferential leap in competency modeling: The effects of task-related information and subject matter expertise. *Personnel Psychology*, 57, 881-904.

Mount, M. K. (1984). Psychometric properties of subordinate ratings of managerial performance. *Personnel Psychology*, 37, 687–702.

Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater and level effects in 360-degree performance ratings. *Personnel Psychology*, 51, 557-576.

Murphy, K., & Cleveland, J. (1995). Understanding performance appraisal: Social, organizational and goal-oriented perspectives. Newbury Park, CA: Sage.

Purkhardt, S.C. and Stockdale, J.E. (1993) "Multidimensional Scaling as a Technique for the Exploration and Description of a Social Representation", pp. 272-297, in G.M. Breakwell and D.V. Canter (eds) *Empirical Approaches to Social Representations*. Oxford: Oxford University Press.

Schippmann, J.S. (1999). The practice of competency modelling. Personnel Psychology, 53, 703-740.

Spencer, L.M., Jr. & Spencer, S.M. (2008). Competence At Work - Models For Superior Performance. New York: Wiley.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 469-477.

Thornton, G.C. (1980). Psychometric Properties of Self-Appraisals of Job Performance. *Personnel Psychology*, 33 (2), 263–271.

Viswesvaran, C., Schmidt, F. & Ones, D. (2005) Is there a general factor in ratings of job performance? a meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, 90:108–131.

Wunderer, R. (2007). Führung und Zusammenarbeit - eine unternehmerische Führungslehre. 7. überarbeitete Auflage. Köln : Luchterhand.

# 5 Measuring overall performance through absolute and relative rating format

## 5.1 Abstract

Performance appraisal is a topic that is of both theoretical and practical importance. As such, it is one of the most researched topics in industrial and organisational psychology. This article examines whether between-individual comparisons differ when using absolute versus relative performance appraisal systems. Secondly the authors examined whether “hard skills” such as professional knowledge and methodological skills or “soft skills” such as personal- and social competence would be a better predictor for overall job performance.

Two different methods have been used to show instrumental independence comparing overall performance scores. Whether we use an absolute rating format on the basis of comparing ratees´ performance with performance standards on a multipoint Likert scale or a relative format comparing the overall performance through direct comparison of ratees using a forced-choice ranking format, we obtain same rankings on overall performance scores. Applying multiple regression the authors could show that four competency categories assessed with an absolute rating format can predict a great deal oft the overall performance level measured on a forced-choice rating format.

Regarding the quality of prediction between “hard and soft skills” we found that methodological competence predicts overall performance level to a large extent whereas social competence is not a valid predictor for overall performance level.

## 5.2 Introduction

Performance appraisal is a key component of human resource (HR) management in most organisations and one of the most critical responsibilities for HR and line managers (Miller & Cardy, 2000). Performance appraisal data are used for many purposes including personnel decisions (e.g., promotions, bonus-pay), employee feedback and development, career development, identification of training needs, and other organisational interventions (e.g., job redesign; Cleveland, Murphy, & Williams, 1989; Latham, Skarlicki, Irvine, & Siegel, 1993; Sulsky & Keown, 1998). Ideally,

performance appraisal results should help managers make informed personnel decisions anprovide information that will best enable them to improve employee performance (DeNisi & Pritchard, 2006; Latham & Mann, 2006).

Not surprisingly, given the important and varied uses for appraisal information, a hallmark of previous research has been to search for ways of maximising the psychometric quality of performance appraisal ratings. Historically (and especially prior to 1980), appraisal research was primarily focused on discovering ways of improving the psychometric quality of performance ratings generated by raters (Landy & Farr, 1980; Sulsky & Keown, 1998). Two popular research streams emerged as a result: (a) rating format research and (b) rater training research. The idea was to develop rating formats and rater training programs that maximise the psychometric quality of performance ratings subjectively generated by performance raters (Murphy & Cleveland, 1995). The successful advancement in this type of research has been confirmed by a very recent study that could deploy the positive impact of frame-of-reference training and the use of descriptively-anchored rating scales on the quality (in terms of rating accuracy and interrater reliability) of ratings in interviews (see Kleinmann et. al., in press).

In this study we focus only on the rater format research rather than the training research. The format of performance evaluation can be classified as either a relative rating system or an absolute rating system (see Cascio, 1998). Relative rating systems ask raters to compare an employee's performance with the performance of other employees. Examples include simple ranking, paired comparisons, and forced distribution. Absolute rating systems ask raters to judge employees' performance on the basis of comparing their performance with performance standards, independent of between-individual comparisons. Examples include critical incidents, behavior checklists, and graphic rating scales. While making judgments about performance under absolute rating systems is supposed to be mainly determined by the discrepancy between the observed or recalled behavior and an absolute performance standard defined in subjective or objective terms (i.e., a behavior-standard discrepancy; Feldman, 1986; Murphy & Cleveland, 1995), a rating score may also reflect between individual comparisons. Research on between-individual contrast effects has shown that evaluation of the current ratee is often contrasted away from the evaluation of the preceding ratees (Maurer & Alexander, 1991; Wexley, Sanders, & Yukl, 1973;

Wexley, Yukl, Kovacs, & Sanders, 1972). In organizations, the influence of between-individual comparisons in evaluations appears to be salient because (a) a typical evaluation often involves multiple ratees (Bernardin & Villanova, 1986) and (b) between individual comparison is a common purpose of performance appraisal (e.g., promotion, bonus allocation, and salary change; Cleveland, Murphy, & Williams, 1989).

Prior conceptual development of performance evaluation has primarily focused on the behavior-standard discrepancy. The cognitive approach to performance appraisal (DeNisi, Cafferty, & Meglino, 1984; Landy & Farr, 1980) stresses performance evaluation as reflecting the "true scores" of performance (Wherry & Bartlett, 1982). Feldman (1986) introduced the concept of the "performance model" to describe the standards used in evaluating performance. In their organizational, goal-based approach to performance appraisal, Murphy and Cleveland (1995) characterized performance judgment as a result of the comparison process that detects the behavior-standard discrepancy. Some other models also proposed that performance judgments are made with reference to implicit or explicit standards (e.g., DeCottis & Petit, 1978; Ilgen, 1983).

Despite the many contributions these models have made to understanding how raters judge and rate individual performance, their fundamental focus is on behavior-standard comparisons. In this article we take on a cognitive perspective positing that in absolute rating formats, commonly used in performance appraisal questionnaires, raters are faced to a cognitive bottleneck. The parallel task of attributing an overall performance ranking among individuals while drawing a specific competency profile over 13 competencies for each employee is a highly demanding task and leads to difficulties in information processing and decision making. We argue that we need to find rating formats that help raters to disentangle overall performance score among ratees from the specific competency profile within an individual. We suggest to doing this two tasks in two different steps. In a first step raters should be able to address their holistic view of their overall performance ranking among their employees. In a second step they should be able to draw a specific competency profile on construct level for each individual. We believe that both tasks can be done using a forced-choice ranking format. We argue that a simple forced-choice tool allowing direct comparisons between individuals is just as effective for measuring overall

performance as the questioning format on a multipoint Likert format over several dimensions.

Our first operationalized hypothesis is, that independently of which rating format we are using, we find the same overall performance score comparing across individuals.

By analyzing and comparing the data of the two different rating formats, an absolute format using behavioural standards and a relative format using forced-choice comparisons, we have also been interested in the question which factors, or competency dimensions, predict the overall performance score. Our second operationalized hypothesis is, that according to literature (Goleman, 1995) personal- and social competence also referred to soft skills, are best to predict overall performance score.

## 5.3 Method

According to our formulation of our hypothesis, we want to find out whether overall performance score does significantly differ when using absolute or relative rating formats.

*Absolute rating format*

In order to measure absolute ratings we used the data of the yearly performance appraisal process of a large Swiss utility company. The performance of employees is measured through superior ratings on Likert based questionnaire forms consisting of 13 competency dimensions. Each dimension contains 4 behaviourally anchored items whereas only the dimension is anchored numerically on a Likert type scale form 1 to 4. The 13 dimensions are segmented into four clusters. The first cluster contains job related professional knowledge and job unrelated organizational knowledge. The second cluster consists of methodological competence, work quality, work quantity, problem solving and creativity. The third cluster refers to personal competence and includes competencies such as achievement orientation, self management, sense of responsibility and learning potential. The forth cluster represents social competences such as communication skills, teamwork and cooperating and customer focus. Data has only been available on cluster level for the 4 different clusters. They represent a mean of the dimensions on each cluster. In order to obtain an overall performance score, values of all 4 clusters have been summed up.

*Relative rating format*

If we are interested in the relative performance, Cascio (1998) suggests using simple ranking, paired comparisons, and forced-choice distribution. For practical reasons we decided to use a computer based forced-choice application to assess overall performance through direct comparison among ratees. Parallel to the absolute rating system, a sample of 32 superiors from different functional areas and hierarchy level with a minimum of five subordinates was asked to do a performance ranking of their subordinates on a two-poled continuum with the following anchors: “exceeds job requirements the most” to “fulfils job requirements the least”. This continuous dimension was numerically anchored from 0 (fulfils the job requirements the least) to 100 (exceeds job requirements the most).

We therefore assessed the performance according to a relative rating system where raters are asked to compare an employee’s performance with the performance of other employees. We have used three measurement points of a time interval of 6 months in order to control for time effect when comparing different rating formats.

We used Pearson’s correlation coefficients to compare overall competence score among ratees for each rater. Since the sampling distribution of Pearson's r is not normally distributed we used "Fisher's 'z'-transformation" that converts Pearson's r's to the normally distributed variable 'z'. Fisher's 'z' was also used for computing confidence intervals on Pearson's correlation.

The table below shows our study design with the two different rating format and three different measurement points:

| | **Relative rating (RR) Forced-choice ranking of employees)** | **Absolute rating (AR) (behavioural standards on Likert scales)** |
|---|---|---|
| **t1: autumn 2007** | x | Data collected |
| **t2: spring 2008** | Data collected | x |
| **t3: autumn 2008** | Data collected | Data collected |

Table 1: study design for comparing two rating formats over time

In order to evaluate the influence of the rating format over time, we calculated Pearson correlation coefficients for the following effects:

Time stability:

- Corr. of AR (t1) and AR (t3)
- Corr. of RR(t2) and RR (t3)

Effect of rating format:

- Corr. of RR (t3) and AR (t3)

Effect of rating format over time:

- Corr. RR (t2) and AB (t1)

In the third measurement condition (t3), which stretched over a time period of 6 weeks in autumn 2008, raters have been asked to use the forced-choice tool to rank the ratees not only on an overall performance level, but also on each competency cluster according the four clusters of the competency framework used in the absolute rating format.

We used a linear multiple regression model to predict overall performance scores comparing the results of both rating formats in three different measurement points.

*Predictor Information*

The 4 competency clusters, as mentioned above, were professional knowledge, methodological competence, personal competence and social competence.

In the absolute rating format, the performance scores on each competency cluster represented the mean score over the dimensions of each competency cluster (min. 0, max. 40 points per cluster).

In the relative rating format the score for each competency cluster has been measured by ranking among ratees on each competency construct on a scale from 0 to 100.

*Criterion information*

The criterion used for this study was an overall measure of job performance. This measure was determined through comparative judgement on a forced-choice distribution. Raters were asked to rank their ratees on a holistic performance level on a continuous scale that was anchored with “exceeds the job requirements the most” (max. 100) and “fulfils the job requirements the least” (min. 0).

Figure 1 illustrates the forced-choice tool to measure overall job performance.

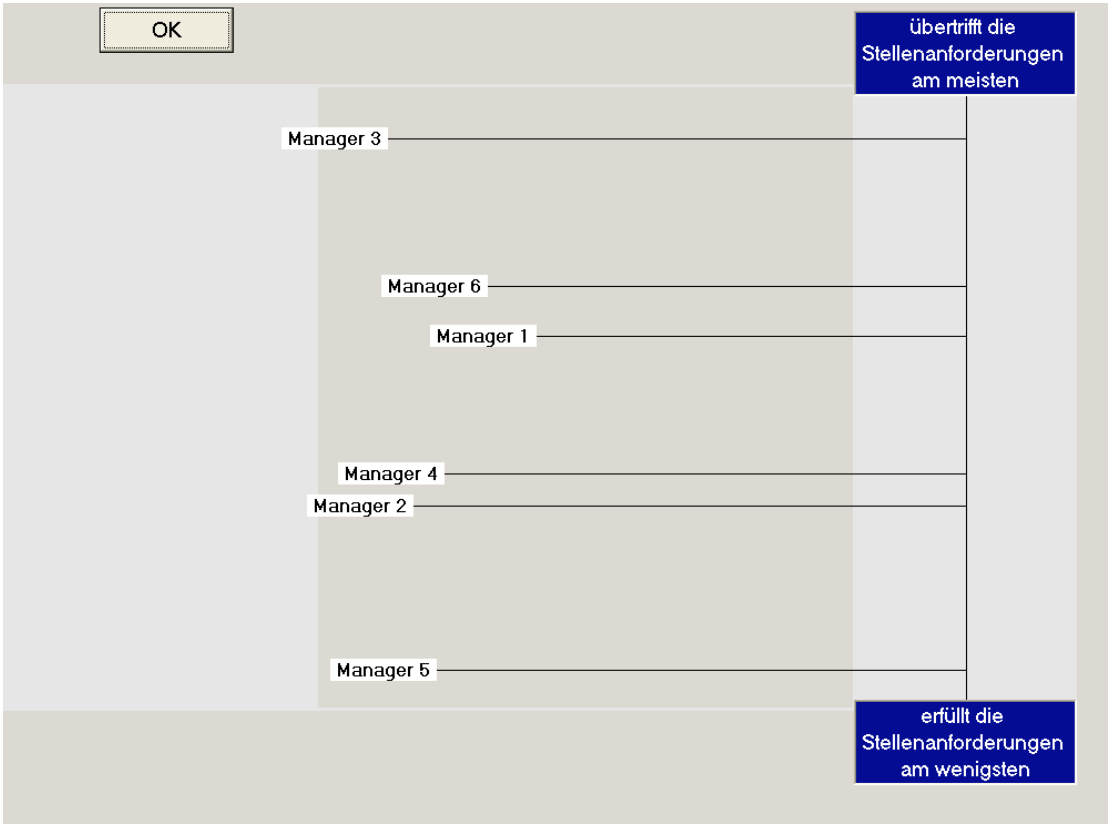



Figure 1: Forced-choice rating tool for assessing overall performance level among ratees.

## 5.4 Results

The results are presented according to our two hypotheses mentioned previously. Our first hypothesis focuses on the question whether the absolute and relative rating format would differ significantly. Secondly we focus on the competency clusters which may predict overall performance level.

*Comparison of absolute and relative rating format*

The correlations between the overall score of absolute rating format (AR) and relative rating format (RR) were calculated separately for each rater's evaluations (for each of n =32 raters, min. 5 and max. 14 ratees).

Table 2 shows all product-moment correlations coefficients between overall performance scores for the different rating formats at different measurement points. We have found high correlations for all measurement conditions when comparing overall score of the two rating formats. This effect is independent over different time intervals of 6 to 12 months.

| Performance Scores (AR, RR, t1-t3) | Pearson R |
|---|---|
| Corr. AR (t1) and AR (t3) | .87*** |
| Corr. RR(t2) and RR (t3) | .91*** |
| Corr. RR (t3) and AR (t3) | .90*** |
| Corr. RR (t2) and AR (t1) | .89*** |

Table 2: Product-Moment Correlations of absolute and relative rating at different measurement points. Note: The correlation coefficients represent the mean of all correlation scores of each rater (n=32). *** p<0.001

*Competency Cluster predicting overall performance score*

As stated in a previous chapter we are interested in the question how different competency clusters, as independent variables, can predict overall performance level. We present results from linear regression models for both types of data: absolute ratings and relative ratings.

Table 3 shows the results of a linear Multiple Regression Analysis using competency constructs as independent variables to explain variance in overall performance. Due to fluctuation and unavailability of data, the sample size varies at different measurement points.

| Predictor | | Overall Performance spring 2008 (t2) | Overall Performance autumn 2008 (t3) |
|---|---|---|---|
| Relative Rating | | | |
| Prof_Comp<br>Meth_Comp<br>Pers_Comp<br>Soc_Comp | 2008 (t2) | X (no data) | X (no data) |
| | 2008 (t3) | R = .68<br>$R^2$ = .45 | R = .89<br>$R^2$ = .78 |
| Absolute Rating | | | |
| Prof_Comp<br>Meth_Comp<br>Pers_Comp<br>Soc_Comp | 2007 (t1) | R = .62<br>$R^2$ = .37 | R = .58<br>$R^2$ = .31 |
| | 2008 (t3) | R = .65<br>$R^2$ = .41 | R = .70<br>$R^2$ = .47 |

Table 3: R and adjusted R Squares depending on measurement point and rating format using competency clusters as predictors. Note: Due to fluctuation and availability of raters the sample size for each measurement condition varies between n = 104 (min.) and n = 190 (max.).After ANOVA all coefficients are highly significant. (*** p<0.001)

All the adjusted $R^2$ are relatively high and indicate that the regression model fit well our data. Looking at the relative rating format we state that the four competency clusters are good predictors for the overall performance level. 78 % of total variance in the overall performance level can be explained through the four predictors. This effect even works when we predict overall performance level in spring 2008 (t2) with performance data on construct level assessed in autumn 2008 (t3), even though the explained variance decreases from 78 % to 45 %. This loss of explained variance is probably due to the change in performance or rating error over time. We will discuss time effects later in this paper.

We are particularly interested in how the amount of explained variance is due to the measurement instrument. If we look at the absolute rating rows we are interested in how well the competency clusters of the absolute rating format can predict the overall performance score that has been assessed through relative rating format. When we use data from the same measurement point, i.e. 2008 (t3), we can explain 47 % of the variance through the four predictors from the absolute rating format measured in autumn 2008 (t3). This effect is relatively stable over time as we can see looking at the results in column spring 2008 (t2); 41% of explained variance, and autumn 2008 (t3), 47% of explained variance. When we compare the effect of the two different instruments at similar measurement points (max. 6 weeks time slot between the two measurements), we find 78 % of variance in the relative rating format and 47% of explained variance in the absolute rating format. This difference is to be attributed to the measurement instrument. In order to examine which competency construct is mainly responsible for predicting overall performance, we list β weights and test for significance whether there is a correlation between predictor and dependant variable.

Table 4 and 5 show the results for the dependant variable measured in spring 2008 (t2) and in autumn 2008 (t3):

| | | Overall Performance spring 2008 (t2) | | |
|---|---|---|---|---|
| Predictor | | β | t | *p* |
| Forced-choice format (t3) | Professional Competence | .229 | 1.901 | .060 |
| | Methodological Competence | .548 | 4.316 | .000*** |
| | Personal Competence | -.113 | -1.022 | .309 |
| | Social Competence | .030 | .319 | .750 |
| Likert format (t1) | Professional Competence | 1.64 | 2.425 | .016* |
| | Methodological Competence | 3.71 | 4.942 | .000*** |
| | Personal Competence | 1.96 | 2.622 | .009** |
| | Social Competence | 0.23 | .34 | .73 |
| Likert format (t3) | Professional Competence | .239 | 3.375 | .001** |
| | Methodological Competence | .290 | 3.988 | .000*** |
| | Personal Competence | .207 | 2.84 | .005** |
| | Social Competence | .106 | 1.596 | .112 |

Table 4: Regression analysis for overall performance level dependant on rating format, measurement point 2

| | | Overall Performance autumn 2008 (t3) | | |
|---|---|---|---|---|
| Predictor | | β | t | *p* |
| Forced-choice format (t3) | Professional Competence | .286 | 3.953 | .000*** |
| | Methodological Competence | .454 | 5.828 | .000*** |
| | Personal Competence | .160 | 2.376 | .019** |
| | Social Competence | .096 | 1.682 | .095 |
| Likert format (t1) | Professional Competence | .260 | 2.575 | .011* |
| | Methodological Competence | .301 | 2.711 | .008** |
| | Personal Competence | .177 | 1.646 | .103 |
| | Social Competence | -.095 | -.996 | .322 |
| Likert format (t3) | Professional Competence | .147 | 1.612 | .110 |
| | Methodological Competence | .384 | 4.033 | .000*** |
| | Personal Competence | .303 | 3.45 | .001** |
| | Social Competence | .029 | .314 | .754 |

Table 5: Regression analysis for overall performance level dependant on rating format, measurement point 3

## 5.5 Discussion

Researchers have looked primarily at two approaches to improve performance appraisal accuracy. Rater training and scale development (Woehr & Huffcutt, 1994). There are at least two important distinctions that can be drawn when examining previous research on alternative rating formats. One concerns the differences between

behaviourally based rating formats and graphic, trait-based rating formats (Aguinis, 2009). The other is between absolute and comparative judgements. Since our absolute rating questionnaires cannot be strictly considered as behaviour anchored rating scales and the forced-choice format at construct level represents a trait based format, we can not draw any conclusions which rating format is more appropriate. Thus, we focus our discussion on the comparison between absolute and comparative judgments.

Landy and Farr (1980) called a moratorium on scale development based on different rating formats because a large body of research had failed to suggest the superiority of any one method over others. However, at the time of the Landy and Farr review, research on scale development was done using rater error measures at the main criteria. Consequently, relatively little scale development research has been conducted in which rating accuracy measures have served as criteria, and this is one reason to lift the moratorium on this line of research (Cardy & Dobbins, 1994). The Landy and Farr (1980) review also marked the beginning of a shift in performance appraisal research towards raters´ cognitive processes (DeNisi, 1996). Despite their call for a moratorium of scale comparison studies, Landy and Farr proposed in their model of the performance rating process that the rating instrument is one of several factors (e.g., the purpose of rating, rater and ratee characteristics) that affect cognitive processes such as observation, storage, and retrieval. This suggestion foreshadowed later calls for an integration of rating format and cognitive research. (e.g., Cardy & Dobbins, 1994; De Nisi. 1996; Murphy & Cleveland, 1995). “Ideally, format should be configured so that the operations required of raters reflect natural cognitive processes leading to efficient and effective processing of performance information” (Bormann, 1991, p.289). Therefore, another reason to rating-scale formats is that they may affect the way in which raters mentally process performance information and could conceivably designed to improve the quality of ratings by using what is known in human cognition (Borman, 1991; Cardy & Dobbins, 1994; DeNisi, 1996; Feldman, 1986; Murphy & Cleveland, 1995). Historically, most relative approaches suffered from serious flaws (e.g., rankings are restricted to an ordinal level of measurement and cannot be meaningfully compared across raters), and this has inhibited research in this area. Among the few innovative approaches to relative rating formats are Miner’s (1988) technique and the Relative Percentile Method (RMP; Goffin, Gellatly, et al., 1996; Wagner & Goffin, 1997). Miner (1988) has described the development and use of a “rated ranking technique” (p. 291) where, similar to the described forced-choice

technique in the present study, raters initially rank their employees on one aspect (dimension) of their job performance. Each employee is then rated by mean of an absolute procedure (i.e., a rater selects a label from outstanding to poor that best describes an individual's performance on a given dimension), and this rating has to be consistent with the initial rank order. Unfortunately, the rated ranking procedure did not demonstrate superiority over straight rankings in terms of reliability and construct validity (Wagner & Goffin, 1997). Another comparative approach is the RPM (Goffin, Gellately, et al., 1996; Wagner & Goffin, 1997). Although the RMP is not yet widely used, it has been relied on to provide criterion data in several publications (e.g., Christiansen, Goffin, Johnston, & Rothstein, 1994; Gellatly, Paunonen, Meyer, Jackson, & Goffin, 1991; Goffin, Rothstein, & Johnson, 1996; Meyer, Paunonen, Gellately, Goffin & Jackson, 1989). RPM is in many ways similar to the forced-choice method used in the present study. RPM is not dependent on an ordinal level of measurement. The RPM also does not impose a particular distribution; Similar to the forced-choice method in this present study, raters of the relative percentile method are free to place ratees anywhere along the 0-100 continuum. Except that no exact ties are allowed. The RPM has typically been applied to measure relatively broad dimensions of job performance rather than critical incidents. Goffin, Gellatly, et. al. found that the RPM showed higher correlations with cognitive ability, personality, and vocational interest measures than did a popular absolute format, the BOS (Lathham & Wexley, 1977), suggesting its greater construct validity. However, Goffin, Gellately, et al. were not able to assess rating accuracy in their field study. A laboratory investigation was conducted to directly assess rating accuracy of the comparative format underlying the RPM (Wagner & Goffin, 1997) where the superiority of relative over an absolute rating format could be shown with Cronbach´s (1955) accuracy components as their dependent variables.

The present study has been conducted in the field using real performance data in the utility branch. We did not use Cronbach´s accuracy components, but we could show that relative rating format leads to equal results in terms of a differentiation of the overall performance level in comparison to absolute rating format. Ratees have reported that it would be a lot easier to do the ratings through direct comparison since they have a precise holistic view of their "excellent performing" and "less performing" employees. This comparison can also be done for each dimension

disentangling overall performance score and competency profile information on construct level. Wagner and Goffin (1997) argued that it might be easier for raters to accurately evaluate performance in comparative (rather than absolute) terms because social comparisons are a natural by-product of uncertain decision-making situations such as performance appraisal. Our argument is more of cognitive nature, stating that step by step comparative judgement on different dimensions is an easier task for raters than questionnaire-based multipoint rating on several dimensions across ratees. The reason for this argument is that the simultaneous task of keeping in mind an overall ranking across individuals while drawing a competency profile on different dimensions for each individual leads to a cognitive overload for most raters. We suggest that performance appraisal should be done in two steps in order to disentangle the two tasks. In a first step raters should be asked to do an overall ranking across individuals, and in a second step to do a ranking on each competency dimension across individuals to give each ratee a specific competency profile. In order to support this practical suggestion with empirical data, we intended to show that the results on overall performance level across individuals are invariant to the measurement instrument being applied. In our findings in table 2 we showed that the overall performance level among individuals in the view of the raters is very stable and does not differ in terms of the rating format or the measurement point. We have found highly significant correlations around $r=.9$ ($p<0.01$) which are rarely found in social sciences and indicate the stability of this effect. Our first hypothesis can therefore be confirmed stating that no matter which rating format we are using, we find the same overall performance scores comparing across individuals. This finding legitimates the application of a forced-choice tool when assessing overall performance level. Moreover, this procedure is inexpensive, easy to carry out and has face validity according to remarks of the raters.

When we recapitulate the findings of our results form the different regression models in table 3, we have some additional evidence that using relative rating format is a veritable alternative to the common used multipoint questionnaires not only on overall performance but also on construct level for different competency clusters. The regressed data from the relative rating format has shown that a four factor model offers an excellent prediction model to estimate overall performance in our data (78% of explained variance).

What is of main interest for our argument of a true alternative using a relative rating format are the findings in the second row of table 3. The four predictors that have been measured through an absolute rating format can explain 47 % of the variance of the overall performance level measured with relative ratings. This effect is relatively stable over different measurement points. The smallest amount of explained variance (30%) we have found when using absolute rating data from autumn 2007 to predict overall performance level of relative format in autumn 2008. The adjusted R square of .30 is still highly significant which means that the predicted scores on the dependant variable based on the values of the four predictors assessed in 2007 correlate with the measured scores of the relative format measured in autumn 2008. This correlation cannot be interpreted by coincidence.

In conclusion we can argue that with the relative rating format we have found much of a better model to predict overall performance through the 4 cluster components than with the absolute rating format. The highest amount of explained variance we could predict with the four components measured through a forced-choice methodology.

Our second hypothesis, which was of minor interest, focuses on the aspects which of the competencies are best predictors for overall performance. The findings of the beta weights and their significance level in table 4 and table 5 are not fully consistent. Only the methodological competence cluster seems to be a solid predictor for overall performance ($p<0.01$). This may also be due to the fact that in the cluster of methodological competence constructs like “quality and quantity of work” are subsumed which are closely related to a general performance score. Professional Competence has also quite high beta weights, but only in 4 out of 6 cases the beta weights are on a significant level. Similar findings we observe for personal competence where beta weights are also in 4 out of 6 cases on a significant level. The pattern changes when we look at the social competence cluster. This cluster sticks out with low beta weights in all measurement conditions. Apparently high scores in social competencies are not related to an overall high performance level. This finding contradicts the famous work of Goleman (1995) stating that emotional intelligence, related to our social competence constructs, is the main driver for overall job performance. We assume that this effect is an artefact of the very technical utility branch where the data has been collected. Mostly professional and methodological

competencies create a halo effect on overall job performance whereas social competence does not significantly correlate with overall job performance.

When we look at the results on table 4 and 5 from a time perspective, we find that in a large time interval of 12 months professional competence is a better predictor than personal competence. This effect reverses when there is a short time slot of a few weeks. Of course these findings have first to be replicated in further research, before plausible reasons for this effect should be established.

Before we discuss the practical implications of these findings, we want to look at the advantages of a forced-choice format over an absolute rating format. One of the main problems of absolute rating formats are the well known rater biases such as leniency, halo and other rater effects. Banks and Murphy (1985) argued that rating performance is a motivated behaviour, and raters may or may not be motivated to rate accurately. Performance appraisals occur in a social context, therefore, it is possible that a rater may decide, for example, to give uniformly high ratings across employees for political reasons (i.e., leniency bias). Similarly, a rater who is concerned with motivating improvement efforts amongst employees may do so by providing them with uniformly average ratings (i.e., range restriction/central tendency error). An important empirical question is whether certain rating formats discourage or otherwise circumvent potential rater tendencies to produce biased ratings—such as rating inflation, truncated range, and so forth? It is the case that particular rating formats, by their construction, can eliminate specific forms of bias. For example, a relative ranking system, requiring raters to make relative comparisons amongst ratees, eliminates the potential for raters to use a truncated range of ratings—simply because no absolute rating score is required. Also, this format eliminates leniency/ severity bias (i.e., a tendency to give overly high or low ratings across ratees) for the same reason (cf. Sulsky & Balzer, 1988).

However, other biases such as personal favouritism can persist when relative assessments are rendered, simply because select ratees are ranked higher (or in some cases lower) than warranted based on their actual performance. There already exists a voluminous literature examining the effect of alternative formats on indices of rater bias such as leniency (e.g., Landy & Farr, 1980). However, conceptual problems with specific operationalisations of bias render this body of research very difficult to interpret (Murphy & Balzer, 1989).

*Practical implications*

Personnel selection and personnel development are important tasks of human resource management. In this study we attempted to link different rating formats to different purposes of performance appraisal. When we are interested in selection of a best performer we look at the overall sum in the performance score across individuals. On the other side, on personal development issues, we are interested in the performance on a construct level within an individual. We posit that performance appraisal should be done in two steps disentangling overall performance across individuals and competency profile on a construct level for each individual. As we have shown with our results, both purposes can be achieved using a forced-choice format which is a more economical and easier method to assess overall performance as well as performance on specific competency constructs. As Wagner and Goffin (1997) suggested, comparative performance appraisals may have the greatest advantage over absolute approaches for purposes such as training needs assessment and employee feedback.

*Limitations*

Some limitations of the study should be considered. The first concerns the homogeneity of our sample and the second, probably the more important one, the inferential leap of job requirement level and performance comparison across individuals.

Employee homogeneity may, however, be common in many work groups, as organizations usually recruit people who are quite similar in their knowledge, skills, and abilities. Future research should examine more heterogeneous work samples across different branches. Especially our effect of methodological superiority over social competencies may be an artefact of the industrial branch and may not be found in more service and social oriented environments.

The second concern refers to comparative rating of ratees on different job requirement levels. The rater is asked to do an inferential leap of different job requirement level and level of performance across different ratees. This may also lead to a cognitive bottleneck. Our results cannot fully prove that this is not the case. We could show that on an overall performance level it does not matter whether the performance level has been measured on an absolute rating format or comparative rating level. But on

construct level we do not know whether the comparative rating lead to valid comparisons across individuals with different job requirement levels. In further research the quality of measurement on construct level needs to be evaluated.

Another concern arising from comparative rating scales is that it may be difficult for a rater to justify or defend a particular ranking, unless the rater has absolute performance information to bolster the comparative assessments. In addition, although some appraisal decisions require that employees simply be ranked or located on a performance curve (e.g., selecting the top employee for promotion), many appraisal decisions will require an absolute assessment (e.g., calibrating bonus pay to the level of performance).

*Conclusion*

We have tried to provide some useful information on the difference between absolute and relative performance measurement and examined how these different types of measurement differ in the level of overall performance. We have found that we can assess overall performance level across different ratees through a lot more efficient method using a forced-choice tool on a continuous scale rather than performance standards questionnaires for each individual. We suggest to practitioners to do performance appraisal in two steps in order to separate the two tasks of differentiating overall performance level and performance on construct level. We suggest further to use comparative rating through a forced-choice rating format on a continuous ranking scale from 0 to 100 with verbal anchors “exceeds job requirements the most” and “fulfils job requirements” the least.

We encourage researchers to do their research also in applied settings. The various socio-political aspects of applied performance appraisal should be of great interest for interpreting empirical research. Murphy and Cleveland (1995) distinguished between performance *judgements*, which are private evaluations of ratee’s performance, and performance *ratings*, which are public statements about ratee’s performance. The latter are more heavily influenced by the rating context, including the social and political environment. The present research included both types by using “real” performance *ratings* in a social and political environment and by measuring performance *judgment* through an additional relative rating format that was assessed in a “private context” for academic purpose. Given its importance to the everyday life of working people, it is safe to assume that research will have to continue to do

research in applied settings in order to legitimate the practical benefit of industrial-organisational psychology research.

## 5.6 References

Aguinis, H. (2009). Performance management. Upper Saddle River, NJ: Pearson Education.

Banks, C. G., & Murphy, K. R. (1985). Toward or narrowing the research practice gap in performance appraisal. *Personnel Psychology,* 38, 335–345.

Bartlett, C. J. (1983). What's the difference between valid and invalid halo? Forced-choice measurement without forcing a choice. *Journal of Applied Psychology*, *57*, 101-109.

Bernardin, H. J., & Villanova, P. (1986). Performance appraisal. In E. Locke (Ed.), *Generalizing from laboratory to field settings* (pp. 189–206). Lexington, MA: Lexington Books.

Cascio, W. F. (1998). Applied psychology in human resource management. Upper Saddle River, NJ: Prentice Hall.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology*, 74, 130–135.

DeCotiis, T., & Petit, A. (1978). The performance appraisal process: Amodel and some testable propositions. *Academy of Management Review*, 3, 635–646.

DeNisi, A. S., & Pritchard, R. D. (2006). Performance appraisal, performance management, and improving individual performance: A motivation framework. *Management and Organization Review*, 2, 253–277.

DeNisi, A. S., Cafferty, T., & Meglino, B. (1984). A cognitive view of performance appraisal process: A model and research propositions. Organizational Behavior and Human Decision Processes, 33, 360–396.

Feldman, J. M. (1986). Instrumentation and training for performance appraisal: A perceptual cognitive viewpoint. In K. Rowland & J. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 4, pp. 101–145). Greenwich, CT: JAI Press.

Ilgen, D. R. (1983). Gender issues in performance appraisal: A discussion of O'Leary and Hansen. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 219–230). Hillsdale, NJ: Erlbaum.

Landy, F. J., & Farr, J. L. (1980). Performance rating. Psychological Bulletin, 87, 72–107.

Latham, G. P., Skarlicki, D., Irvine, D., & Siegel, J. P. (1993). The increasing importance of performance appraisals to employee effectiveness in organizational settings in North America. *International Review of Industrial and Organizational Psychology,* 8, 87-131.

Melchers, K. G., Lienhardt, N., von Aarburg, M., & Kleinmann, M. (in press). Is more structure always better? An evaluation of the effects of rater training and de scriptively anchored rating scales on rating accuracy in a structured interview. *Personnel Psychology.*

Latham, G. P., & Mann, S. (2006). Advances in the science of performance appraisal: Implications for practice. *International Journal of Organizational and Industrial Psychology*, 21, 295–337.

Maurer, T. J., & Alexander, R. A. (1991). Contrast effects in behavioural measurement: An investigation of alternative process explanations. *Journal of Applied Psychology*, 76, 3–10.

Murphy, K. R., & Cleveland, J. N. (1995). Understanding performance appraisal. Thousand Oaks, CA: Sage.

Sulsky, L. M., & Keown, J. L. (1998). Performance appraisal in the changing world of work: Implications for the meaning and measurement of work performance. *Canadian Psychology*, 39, 52–59.

Miller, J. S., & Cardy, R. L. (2000). Self-monitoring and performance appraisal: Rating outcomes in project teams. *Journal of Organizational Behavior*, 21, 609–626.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, 73, 497–506.

Wagner, S. H., & Goffin, R. D. (1997). Differences in accuracy of absolute and comparative performance appraisal methods. *Organizational Behavior and Human Decision Processes*, 70, 95–103.

Wexley, K. N., Yukl, G. A., Kovacs, S. Z., & Sanders, R. E. (1972). Importance of contrast effects in employment interviews. *Journal of Applied Psychology*, 56, 45–48.

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology*, 35, 521–555.

# 6 Relationale Kompetenzmessung und deren Visualisierung

## 6.1 Einleitung

Der Kompetenzbegriff ist sowohl in der Management Literatur als auch im betrieblichen Alltag nicht mehr wegzudenken. Computer- und Medienkompetenz (Gapski 2001) werden erwartet, Führungs- (Jetter et al. 2001) und Coachingkompetenz (Bayer 1995) gefordert, Organisations- (Thom & Zaugg 2001) und Selbstorganisationskompetenz (North 1999) gefördert. Kompetenzmanagement (Probst et al. 2000) ergänzt das schon gängige Wissensmanagement (Probst et al. 1999). Der mit interkultureller Kompetenz (Kalpaka 1998) ausgestattete Kompetenzmensch wird zum höchsten Ziel lebenslangen Lernens (Wildmann 2001). Davon ausgehend muss verwundern, wie wenig klar „Kompetenz“ gegenwärtig begrifflich gefasst und messend zugänglich gemacht werden kann.

Die deutsche Kompetenzforschung hat zahlreiche Beiträge zum Thema Kompetenzmessung hervorgebracht. Einen umfassenden und recht aktuellen Überblick über die verschiedenen in diesem Rahmen entwickelten Kompetenzmessverfahren liefern Erpenbeck und Von Rosenstiel (2007). Die meisten dort beschriebenen Verfahren gehen nach dem Prinzip der gängigen persönlichkeitspsychologischen Testkonstruktion vor oder bewegen sich in der Tradition der aus Arbeits- und Organisationspsychologie bekannten und von Schuler & Funke (1995) beschriebenen Simulationsverfahren, welche in der Praxis in Assessment Centern ihren Niederschlag finden.

Im Folgenden stellen wir einen etwas anderen Zugang zur Kompetenzmessung vor. Er bietet vor allem neue Anwendungsmöglichkeiten bei der Visualisierung von Kompetenzprofilen von Führungskräften. Basierend auf dem von Wunderer (2007) postulierten Konzept des Mitunternehmertums wollen wir illustrieren, wie Kompetenzprofile von Führungskräften auf die drei Schlüsselkompetenzen Gestaltungskompetenz, Soziale Kompetenz und Umsetzungskompetenz reduziert werden können und damit ein schlankes und effizientes Modell bietet, welches Führungskräfte beziehungsweise deren unterschiedliche Kompetenz-Ausprägung in den drei Schlüsselkompetenzen auf einen Blick erfasst. Daraus lassen sich

praxistaugliche Massnahmen ableiten wie zum Beispiel stellenspezifische Nachfolgeplanung oder mitarbeiterorientierte Karriereplanung.

## 6.2 Methodik

Die vorliegende Studie ist methodisch in zwei Teile aufgeteilt. Der erste Teil zeigt auf, wie sich anhand eines in der Praxis angewendeten Kompetenzmodells durch ein Forced-Choice-Verfahren Kompetenzprofile von Führungskräften auf mittlerer und oberer Managementstufe bilden lassen. Die durch dieses Verfahren ermittelten Kompetenzprofile lassen sich dann mit dem dimensionsreduzierenden Verfahren der Nonmetrischen Multidimensionalen Skalierung (NMDS) miteinander in Beziehung setzen. Das in dieser Studie angewendete Verfahren basiert auf einem nach Läge et al. (2005) konstruierten robusten Algorithmus, welches Ähnlichkeitsbeziehungen zwischen Objekten, in diesem Fall Kompetenzprofilen, mit einem geringst möglichen Fehlermass optimal in einen zweidimensionalen Raum einpasst. Ähnliche Profile werden dabei nahe beieinander abgebildet, während unähnliche Profile weit voneinander entfernt zu liegen kommen. Der Vorteil in der Darstellung der NMDS liegt darin, dass durch die Visualisierung der Daten die Nachbarschaftsbeziehungen zwischen den verschiedenen Objekten auf einen Blick ersichtlich werden. Anders als bei den gängigen dimensionsreduzierenden Analyseverfahren wie z.B. bei der Faktorenanalyse wird bei der NMDS die gesamte Varianz gleichmässig berücksichtigt, ohne dass die durch die Hauptkomponenten nicht erklärte Varianz systematisch verloren geht.

In einem zweiten Teil werden die 75 Kompetenzbegriffe des hierarchischen Kompetenz-Modells durch ein Expertenurteil in die drei von Wunderer postulierten Schlüsselkompetenzen Gestaltungskompetenz, Soziale Kompetenz und Umsetzungskompetenz eingestuft. Durch diese Einstufung werden die Kompetenzprofile auf drei Dimensionen reduziert. Die Kompetenzprofile können wiederum mittels NMDS skaliert werden, wobei das Ähnlichkeitsmass in diesem Fall nur auf drei Werten beruht. Mittels des anschliessenden Property Fittings[6] der drei Kompetenzdimensionen kann dann gezeigt werden, wie sich die Führungskräfte

[6] Skalen und Rangreihen können mittels einer Multiplen Regression in die Struktur hineingelegt werden und ermöglichen so eine dimensionale Interpretation der Karte. Dabei wird ein Regressionskoeffizient als Gütekriterium für die jeweilige Dimension angegeben.

anhand eines schlanken Kompetenzmodells von drei Schlüsselkompetenzen beschreiben lassen.

*Bildung von Kompetenzprofilen mittels Forced-Choice-Verfahren*

Die Daten der vorliegenden Studie wurden mittels eines in der Praxis entwickelten Kompetenzmodells in einem grösseren Schweizerischen Energieunternehmen erhoben. Bevor auf das Forced-Choice-Messverfahren eingegangen wird, soll an dieser Stelle das zu Grunde liegende Kompetenzmodell beschrieben werden.

Basierend auf der Analyse der gängigen Kompetenzmodelle[7] in der Praxis und den Erkenntnissen wissenschaftlicher Kompetenzforschung[8], wurde für die empirische Untersuchung ein Kompetenzmodell postuliert, welches hierarchisch in 4 Kompetenzfelder, 15 Kompetenzdimensionen und 60 Kompetenzfacetten aufgebaut ist.

| **Fokus - Erfolgsstrategien entwickeln** | | **Leistungsverhalten - Initative ergreifen und in hoher Qualität umsetzen** | |
|---|---|---|---|
| **Unternehmerisches Denken und Handeln** | Marktorientierung<br>Chancen-/Risikobewusstsein<br>vernetztes Denken<br>Kosten-/Nutzen Denken | **Leistungsmotivation** | Ambition<br>Initiative<br>Verantwortungsübernahme<br>Engagement |
| **Problemlösungsfähigkeit** | Analytische Fähigkeiten<br>system.-methodisch. Vorgehen<br>Schlussfolgerndes Denken<br>Urteilsvermögen | **Selbstmanagement** | Umgang mit eigenen Ressourcen<br>Selbstreflexionsfähigkeit<br>Belastbarkeit<br>Überblick bewahren |
| **Planungs-und Organisationsfähigkeit** | Projektmanagement<br>Prozessdenken<br>Finanz- und Ressourcenmangt<br>Prioritäten setzen | **Umsetzungsorientierung** | Entscheidungsfähigkeit<br>Ergebnisorientierung<br>Durchsetzungsvermögen<br>Beharrlichkeit |
| **Offenheit für Neues** | Antizipationsfähigkeit<br>Flexibilität<br>Lernfähigkeit<br>Innovationskraft | **Qualitätsbewusstsein** | Verbesserungsstreben<br>Verlässlichkeit<br>Genauigkeit<br>Sicherheitsbewusstsein |
| **Kundenorientierung** | Kundenbedürfnisse erkennen<br>Beratungsfähigkeit<br>Kundenfreundlichkeit<br>Netzwerke pflegen | | |
| **Soziale Kompetenz - Partner gewinnen** | | **Leadership - Team führen** | |
| **Kommunikationsfähigkeit** | Information & Rückmeldung<br>Verhandlungsfähigkeit<br>Gesprächsführung<br>Präsentationsfähigkeit | **Andere Inspirieren** | Sinn vermitteln<br>Veränderungsbereitschaft<br>Integrität<br>Begeisterungsfähigkeit |
| **Zusammenarbeit / Teamfähigkeit** | Teamorientierung<br>Hilfsbereitschaft<br>Integrationsfähigkeit<br>Kooperationsfähigkeit | **Führungsverhalten** | Zielorientiertes Führen<br>Fähigkeit zur Delegation<br>Teamleistung fördern<br>Zielerreichung überprüfen |
| **Konfliktlösungsfähigkeit** | Zwischenmenschliches Feingespür<br>Perspektivenübernahme<br>Konfliktfreudigkeit<br>Umgang mit Kritik | **Mitarbeiterförderung** | Vertrauen schaffen<br>Feedback geben<br>Leistung einfordern<br>Andere fördern |

Abbildung 1: Hermeneutisches Kompetenzmodell als Datengrundlage der empirischen Studie

[7] Bei der Analyse wurden branchenfremde Modelle von Siemens, ABB, IBM, KMPG, Daimler Chrysler, Credit Suisse sowie das Schweizerische Militär berücksichtigt, während bei der branchenspezifischen Analyse die Kompetenzmodelle der CKW AG, Axpo Informatik AG, NOK AG, EGL AG und Stadtwerke Münster einbezogen wurden.

[8] Erpenbeck & Von Rosenstiel (2003), Spencer & Spencer (2008) Boyatzis (2006), Schippmann (2000); Dulewicz & Young (2008).

Die Zuordnung der Kompetenzdimensionen zu vier Kompetenzfeldern ist stark an die Modelle in der Praxis angelehnt. Gemäss dem Kompetenzraster von Erpenbeck und Heyse (1999) wurden bei der Auswahl berufsrelevanter Kompetenzfacetten Personale Kompetenzen, Aktivitäts- und Handlungskompetenzen, Sozial-Kommunikative Kompetenzen sowie Fach- und Methodenkompetenzen berücksichtigt.

Für die vorliegende Studie wurde dann ein Messverfahren entwickelt, welches lediglich das Profil und nicht die Profilhöhe, d.h. den absoluten Ausprägungsgrad berufsrelevanter Kompetenzen, abfragt. Das Verfahren sieht vor, die einzelnen Kompetenzbegriffe und deren Definitionen in einem anwenderfreundlichen Online-Tool in zwei Schritten in eine Rangreihe zu bringen. Durch dieses Verfahren werden Kompetenzprofile gebildet, welche die relativen Stärken und Schwächen einer Person abbildet. Führungskräfte aus dem mittleren und oberen Managements aus unterschiedlichen Fachbereichen wurden mittels dieses onlinebasierten Kompetenzerhebungverfahrens befragt.

In einem ersten Schritt wurden die Führungskräfte gebeten, die 15 Kompetenzdimensionen in drei Kategorien zu teilen, je nach Selbsteinschätzung des Ausprägungsgrades, von „weniger stark ausgeprägt“ bis „sehr stark ausgeprägt“ (Vgl. Abbildung 2).

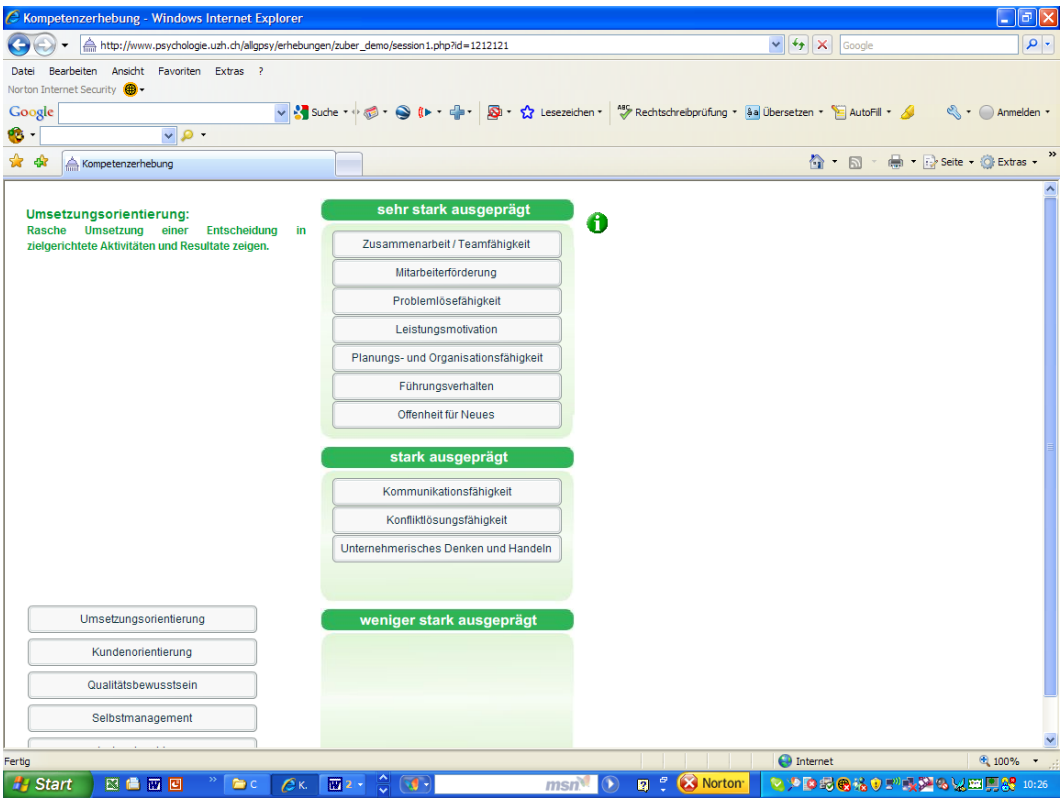



Abbildung 2: Kompetenzprofilbildung durch onlinebasiertes Forced-Choice-Tool, 1. Schritt

In einem zweiten Schritt (vgl. Abbildung 3) wurden die Versuchspersonen gebeten, die Kompetenzen pro Kategorie in eine Rangreihe zu bringen. Durch dieses Verfahren wurde pro Versuchsperson ein auf Rangplatz basierendes Kompetenzprofil erstellt. Somit wurde die Datenbasis für Rangkorrelationen zwischen Kompetenzprofilen geschaffen. Die Rangkorrelationen beruhen jeweils auf 15 Werten pro Person.

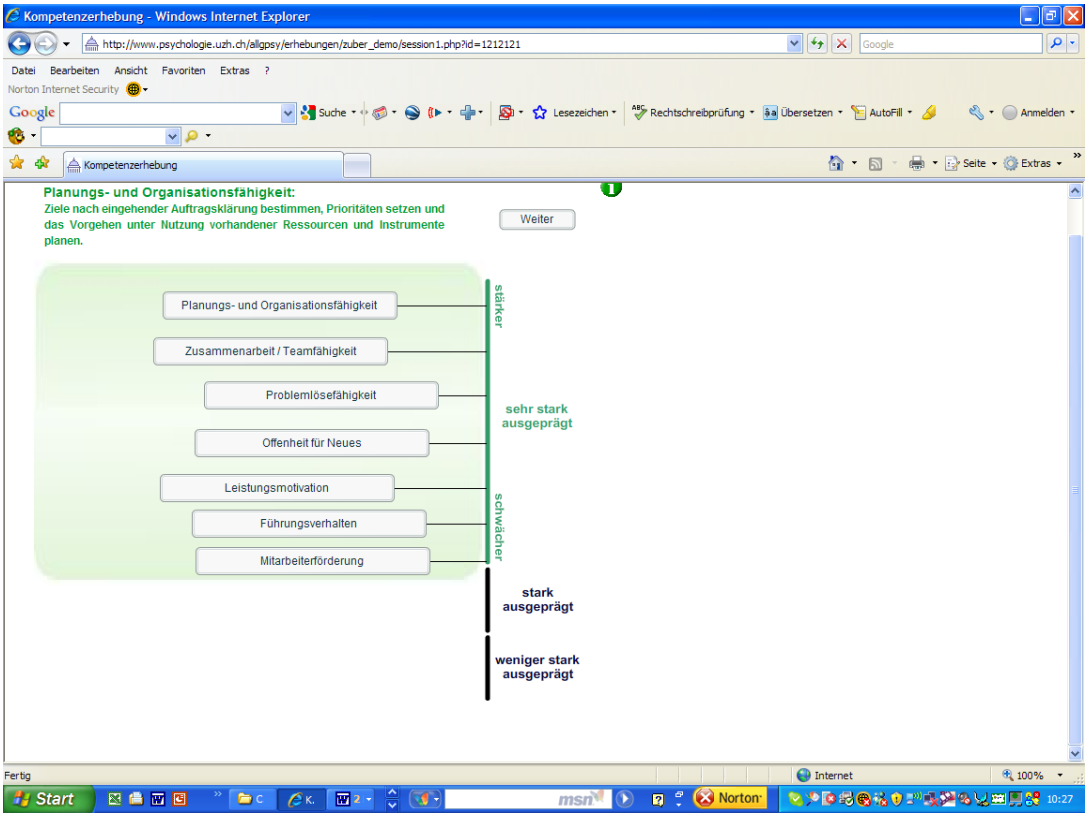



Abbildung 3: Kompetenzprofilbildung durch onlinebasiertes Forced-Choice-Tool, 2. Schritt

Basierend auf Selbst- und Fremdeinschätzungen zu zwei Erhebungszeitpunkten im Abstand von 2 Wochen wurden insgesamt 41 Kompetenzprofile erhoben, wobei 28 Versuchspersonen die Messbedingungen[9] erfüllten und somit in die Auswertung flossen. Alle 75 Kompetenzbegriffe (15 Kompetenzdimensionen +60 Kompetenzfacetten) wurden in der Selbsteinschätzung gemäss dem oben beschriebenen Verfahren abgefragt.

[9] Als Messbedingung wurden folgende drei Kriterien herangezogen: 1. Teilnahme an beiden Erhebungszeitpunkten, 2. Stabilität des Selbstbildes (Korrelation zwischen Selbstbildern zu Messzeitpunkt 1 und Messzeitpunkt von mind. r=.5), 3. Dauer der Erhebung: 3 Personen wurden aufgrund unrealistisch kurzer Erhebungsdauer (weniger als 1 Minute) aus der Untersuchung entfernt.

Die Fremdeinschätzungen erfolgten nur auf den 15 Kompetenzdimensionen und sind nicht Gegenstand der vorliegenden Studie.

*Einordnung der Kompetenzbegriffe in eine Struktur bestehend aus 3 Kategorien*

Ziel dieser Studie ist die einfache Darstellung von Kompetenzprofilen mittels einer geringen Anzahl an Kompetenzdimensionen.

Im Konzept des Mitunternehmertums unterscheidet Wunderer (2007) drei notwendige und hinreichende Schlüsselkompetenzen zur erfolgreichen Bewältigung der Führungsaufgaben und Erreichung der Organisationsziele: Gestaltungskompetenz, Umsetzungskompetenz und Sozialkompetenz.

Die vorwiegend kognitive Gestaltungskompetenz wird bei Wunderer (2007) definiert als eine Begabung und Motivation zu innovativ-gestalterischer Aktivität im Dienste der Organisationsziele bzw. –Strategie. Die aktionale Umsetzungskompetenz bezieht sich gemäss Wunderer auf die Fähigkeit und Bereitschaft zur effizienten Verwirklichung und Implementierung innovativer Problemlösungen. Als drittes Element erfolgreichen Mitunternehmentums beschreibt Wunderer die Sozialkompetenz als Kooperations- und Integrationsfähigkeit und –Motivation, die zur selbstorganisierten und zugleich kooperativen Verwirklichung von innovativen Ideen im Team oder über Abteilungsgrenzen hinweg dient.

In der vorliegenden Studie wurde untersucht, in wiefern die Daten basierend auf einem umfangreichen Kompetenzmodell von 15 Kompetenzdimensionen und 60 Kompetenzfacetten sich in die Terminologie der drei Schlüsselkompetenzen transferieren lassen und in einem zweidimensionalen Modell abgebildet werden können. Zu diesem Zweck wurden die einzelnen Kompetenzbegriffe durch ein Expertenurteil von Wunderer den einzelnen Schlüsselkompetenzen zugeordnet. Abbildung 5 zeigt die Instruktion dieser Erhebung.

Universität St.Gallen — **Validierungsstudie Kompetenzstruktur** — Universität Zürich

**Instruktion:**
In der Folge sind 75 Kompetenzbegriffe mit den jeweiligen Beschreibungen aufgelistet. Ziel der vorliegenden Erhebung ist die Einordnung der Kompetenzbegriffe in die 3 dimensionale Kompetenzstruktur "Gestaltungskompetenz, Umsetzungskompetenz und soziale Kompetenz".
Bitte vergeben Sie in der Spalte "rating" gemäss unten stehender Legende die entsprechnede Codierung.

**1 = Gestaltungskompetenz**
**2 = Umsetzungskompetenz**
**3 = Soziale Kompetenz**

4 = keine klare Zuordnung möglich (Bitte fügen Sie in der Spalte "Bemerkungen" diese Codierung ein, falls es für Sie keine eindeutige Zuordnung der jeweiligen Kompetenz gibt.

| Kompetenzbegriff | Beschreibungen | Rating | Bemerkung: |
|---|---|---|---|
| Unternehmerisches Denken und Handeln | Chancen und Risiken erkennen, Vorgehensweisen zur Effektivitäts- und Effizienzsteigerung definieren und Mehrwert schaffen. | | |
| Problemlösefähigkeit | Probleme gründlich analysieren, sie systematisch in bearbeitbare Teilaufgaben zerlegen und zielführende Lösungen finden. | | |
| Planungs- und Organisationsfähigkeit | Ziele nach eingehender Auftragsklärung bestimmen, Prioritäten setzen und das Vorgehen unter Nutzung vorhandener Ressourcen und Instrumente planen. | | |
| Offenheit für Neues | Neue Entwicklungen positiv aufnehmen, sich lernfähig zeigen und innovative Ideen für die Zukunft entwickeln. | | |
| Kundenorientierung | Tragfähige Arbeitsbeziehungen zu internen und externen Kunden knüpfen, auf deren Bedürfnisse eingehen und passende Kundenlösungen entwickeln. | | |
| Kommunikationsfähigkeit | Informationen klar und präzis vermitteln, aktiv zuhören und andere durch nachvollziehbare Argumente überzeugen. | | |
| Zusammenarbeit / Teamfähigkeit | Sich in Gruppen einfügen, kooperieren und einen konstruktiven Beitrag zur gemeinsamen Zielerreichung leisten. | | |
| Konfliktlösungsfähigkeit | Meinungsverschiedenheiten / Konflikte direkt ansprechen, eine selbstkritische Grundhaltung haben und aktiv zu einer Lösung beitragen. | | |

Abbildung 5: Instruktion Validierungsstudie mit drei Schlüsselkompetenzen nach Wunderer

Durch eine Reihenfolge der entsprechenden Zuordnung von Kompetenzbegriff und Schlüsselkompetenz wurde eine entsprechende Gewichtung vorgenommen, in wie fern der einzelne Kompetenzbegriff ein guter Vertreter der jeweiligen Schlüsselkompetenz ist. Die Reihenfolge der Zuordnungen zu den drei Schlüsselkompetenzen wurde mit den Gewichten 1.5 (Rang 1), 1 (Rang 2) und 0.5 (Rang 3) berücksichtigt. Durch Multiplikation der Rohwerte aus den Kompetenzprofilen und des Gewichtungsfaktors gemäss der Zuordnung von Wunderer konnte pro Versuchsperson ein Kompetenzwert pro Schlüsselkompetenz ermittelt werden. Abbildung 6 verdeutlicht die Rechenschritte zur Ermittlung des Kompetenzprofils basierend auf den drei Schlüsselkompetenzen von Wunderer:

**Vorgehen Regressionsanalyse "Kompetenzdimenisionen": Modell Wunderer**

**Schritt 1**

| | Unt. D&H | Problfäh. | P&O-Fäh | Off_f-Neu | Kundorient. | etc. |
|---|---|---|---|---|---|---|
| VP 1 | 8.5 | 13 | 14.5 | 2.5 | 1.5 | |
| VP 2 | 13.5 | 7.5 | 9.5 | 8 | 13.5 | |
| VP 3 | 1.5 | 1.5 | 13 | 5 | 3 | |
| VP 4 | 10.5 | 6.5 | 2.5 | 7.5 | 11 | |
| VP 5 | 2.5 | 13.5 | 9 | 1 | 13.5 | |
| VP 6 | 12 | 10 | 15 | 1.5 | 6.5 | |
| VP 7 | 12.5 | 1.5 | 6 | 12 | 11 | |
| VP 8 | 4 | 10 | 14 | 2.5 | 6 | |
| VP 9 | 9 | 7 | 13.5 | 6.5 | 8 | |
| VP 10 | 12 | 5.5 | 8.5 | 10 | 4.5 | |
| etc. | | | | | | |

**Schritt 2**

| | **VP 1** | Gewichte | GK | UK | SK | **GL** | **UK** | **SZ** |
|---|---|---|---|---|---|---|---|---|
| Unt. D&H_K | 8.5 | 1,3,2 | 1.5 | 0.5 | 1 | 12.75 | 4.25 | 8.5 |
| Problfäh._K | 13 | 1 | 3 | 0 | 0 | 39 | 0 | 0 |
| P&O-Fäh_K | 14.5 | 1 | 3 | 0 | 0 | 43.5 | 0 | 0 |
| Off_f-Neu_K | 2.5 | 1 | 3 | 0 | 0 | 7.5 | 0 | 0 |
| Kund_or_K | 1.5 | 3,2,1 | 1.5 | 1 | 0.5 | 2.25 | 1.5 | 0.75 |
| Kommfäh_K | 4 | 3,1 | 1 | 0 | 2 | 4 | 0 | 8 |
| Teamfäh_K | 2.5 | 3 | 0 | 0 | 3 | 0 | 0 | 7.5 |
| Konflikt_K | 8.5 | 3 | 0 | 0 | 3 | 0 | 0 | 25.5 |
| Leistmot_K | 5 | 3,2 | 0 | 1 | 2 | 0 | 5 | 10 |
| | | | | | | | | |
| **Summe** | | | 13 | 2.5 | 11.5 | 8.38 | 4.30 | 5.24 |

**Schritt 3**

| | GK | UK | SK |
|---|---|---|---|
| VP01 | 8.38 | 4.30 | 5.24 |
| VP02 | 9.31 | 7.29 | 7.39 |
| VP03 | 7.00 | 8.02 | 8.80 |
| VP04 | 8.34 | 6.87 | 8.45 |
| VP05 | 8.43 | 7.28 | 8.11 |
| VP06 | 9.06 | 8.38 | 6.90 |
| VP07 | 7.89 | 8.27 | 7.91 |
| VP08 | 7.65 | 7.69 | 8.48 |
| VP09 | 8.70 | 7.05 | 8.04 |
| VP10 | 10.14 | 7.48 | 6.60 |

**Schritt 1:**

Kompetenzdimensionen und - Facetten

VP = Versuchspersonen (Führungskräfte)

Rohwerte gemäss Forced Choice Profile

**Schritt 2:**
Gewichtungsmodell gemäss Einstufung der Kompetenzbegriffe zu den 3 Schlüsselkompetenzen.
Beispiel „Unternehmerisches Denken und Handeln
1 = Gestaltungskompetenz (GK), Gew.fakt. 1.5
3 = Soziale Kompetenz (SK), Gew.fakt. 1
2 = Umsetzungskompetenz (UK),Gew.fakt. 0.5

Berechnung des Kompetenzprofils reduziert auf die drei Schlüsselkompetenzen:
Multiplikation des Rohwertes mit dem Gewichtungsfaktor gemäss der Zuordnung von Wunderer.

**Schritt 3:**
Berechnung des Kompetenzwertes pro Versuchsperson:
Summe Produkte (Rohwert x Gewichtungsfaktor) über alle Kompetenzbegriffe geteilt durch die Summe der Gewichtungsfaktoren.

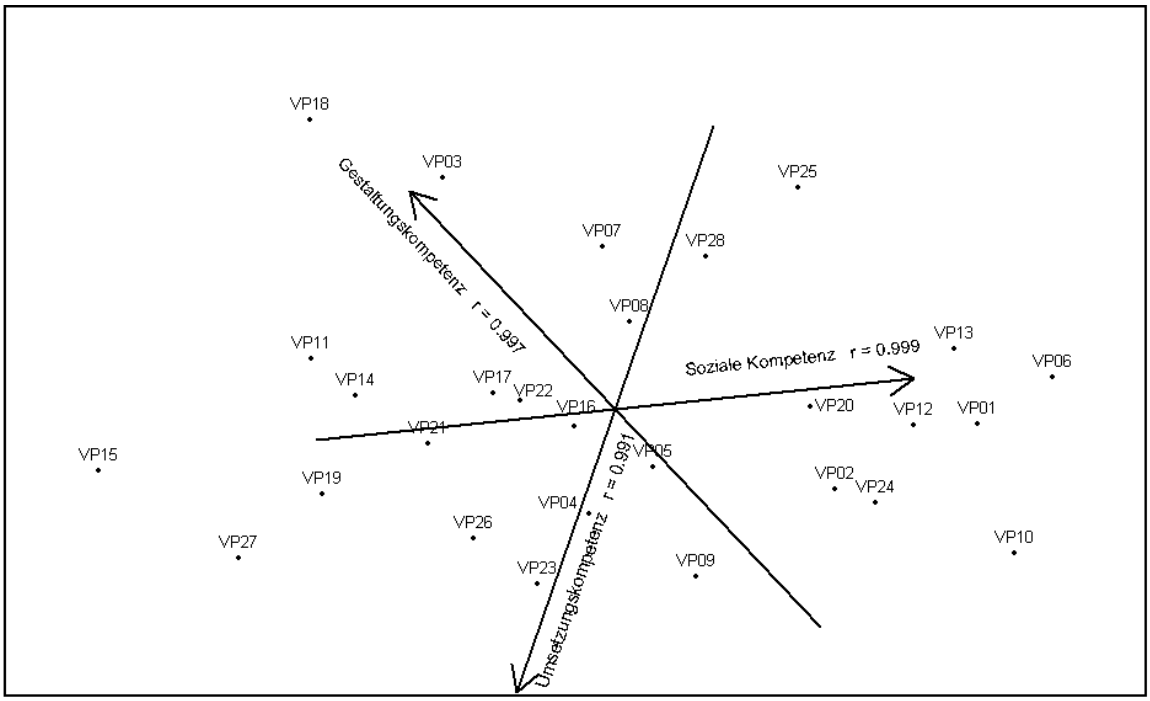
Abbildung 6: Illustration der Berechnung der drei Kompetenzdimensionen nach Gewichtung Wunderer

Durch dieses Gewichtungsverfahren gemäss der Zuordnung von Kompetenzbegriffen zu den drei Schlüsselkompetenzen konnte pro Versuchsperson das Kompetenzprofil von 15 Kompetenzdimensionen auf drei Kompetenzdimensionen reduziert werden.
Die neu ermittelten Kompetenzprofile, welche nun auf drei Werten basieren, können mittels NMDS skaliert und in einen zweidimensionalen euklidischen Raum abgebildet werden. Durch Property Fitting, ein multiples Regressionsverfahren, werden die drei Achsen Gestaltungskompetenz, Umsetzungskompetenz und Soziale Kompetenz in diese Karte gelegt. Dies zeigt auf, inwiefern sich die gewichteten Kompetenzprofile durch die 3 Achsen von Wunderer beschreiben lassen.

## 6.3 Ergebnisse

In dieser Studie wurden mittels eines Forced-Choiced-Verfahrens 28 Kompetenzprofile von Führungskräften bestehend aus 75 Kompetenzbegriffen erhoben.

Die unten abgebildete euklidische Karte zeigt eine „Managerlandschaft“ eines grossen Energieunternehmens mit den Ähnlichkeitsbeziehungen der Kompetenzprofile untereinander.




Stress NMDS: 0.037

Abbildung 7: NMDS und Property Fitting der 3 Schlüsselkompetenzen nach Wunderer

Der tiefe Stresswert von 0.037, das Gütemass der NMDS, welches die Passung der Objekte im zweidimensionalen Raum angibt, deutet darauf hin, dass das Modell einen grossen Teil der Varianz erklärt und die relationalen Ähnlichkeitsbeziehungen zwischen den Versuchpersonen basierend auf den 3 Werten der Schlüsselkompetenzen nahezu optimal abbildet.

Mittels der multiplen Regression wurde im Property Fitting die einzelnen Kompetenzachsen nach Wunderer in die Karte hineingelegt. Die genaue Lage der Achse wird durch die lineare Regression nach dem Prinzip der kleinsten abweichenden Quadrate bestimmt. Die Regressionsgeraden stehen ungefähr im 60°-Winkel zueinander und die hohen Korrelationskoeffizienten von 0.99 sprechen für die optimale Passung der Dimensionen in die Karte.

Die euklidische Karte lässt demnach die Profile der einzelnen Führungskräfte in Bezug auf die Ausprägung in einer der drei Schlüsselkompetenzen relativ exakt interpretieren. Links oben in der Karte befinden sich diejenigen Führungskräfte, welche in der Gestaltungskompetenz ihre Stärken haben, auf der rechten Seite der Karte liegen Führungskräfte, welche insbesondere in den sozialen Kompetenzen ihre Stärken haben, während im unteren Teil der Karte, die Führungskräfte liegen, die vor allem in der Terminologie von Wunderer umsetzungsstark sind. In der Mitte der Karte sind Führungskräfte abgebildet, welche sich nicht durch ein klares Profil in Bezug auf die drei Schlüsselkompetenzen von Wunderer beschreiben lassen, sondern sich vielmehr durch „ein bisschen von allem“ auszeichnen.

## 6.4 Diskussion

Die vorliegende Studie hatte zum Ziel zu zeigen, wie sich Stärken- Schwächenprofile von Führungskräften basierend auf einer umfassenden Auswahl berufsrelevanten Kompetenzen relativ einfach messen und darstellen lassen. Das Forced-Choice-Verfahren mittels eines einfach handhabbaren Online-Tools, welche die Methode der Idealpunktverfahren aus der Marktforschung auf die Kompetenzmessung überträgt, konnte als ein praktikables Verfahren vorgestellt werden, welches die Beurteilung von Kompetenzen mittels Selbst- und Fremdbeurteilung zulässt. Dabei konnte weiter gezeigt werden, dass sich die drei Schlüsselkompetenzen von Wunderer sehr gut eignen, um die mittels einer relativ hohen Anzahl an Kompetenzen erhobenen Kompetenzprofile in einer zweidimensionalen Karte abzubilden und mittels Property Fittings auf einen Blick zu interpretieren. Die Darstellung der Kompetenzprofile basierend auf deren Ähnlichkeitsbeziehungen in einer euklidischen Karte ist eine in der Arbeits- und Organisationspsychologie noch nicht angetroffene Methode der Visualisierung von Kompetenzdaten. Die Abbildung mehrerer Personen lässt auf einen Blick Gemeinsamkeiten und Unterschiede in den Kompetenzprofilen der Führungskräfte erkennen. Durch die Methode des Property Fitting können beliebige Kompetenzmodelle in unterschiedlicher Dimensionalität in eine euklidische Karte gelegt werden, was eine einfache und schnelle Interpretation der Karte in der Kompetenzterminologie des jeweiligen Unternehmens zulässt.

Im Zusammenhang mit der praktischen Anwendung von euklidischen Karten für personale Zwecke ist die Verbindung von Job- und Personenprofilen innerhalb einer

zweidimensionalen euklidischen Karte zu erwähnen. Dabei werden sowohl Anforderungsprofile als auch Kompetenzprofile skaliert und in einen gemeinsamen Raum abgebildet. Dank dieser Methode ergeben sich interessante Ansätze für die gängigen Fragen der Laufbahn- bzw. Nachfolgeplanung. Welche Stelle passt am besten zu einem bestimmten Kandidaten bzw. welche Kandidaten passen am besten auf eine bestimmte Stelle? Abbildung 8 illustriert schematisch die praktische Anwendung euklidischer Karten in der Personalarbeit.

**Beispiel Laufbahnplanung:** **Beispiel Nachfolgeplanung:**

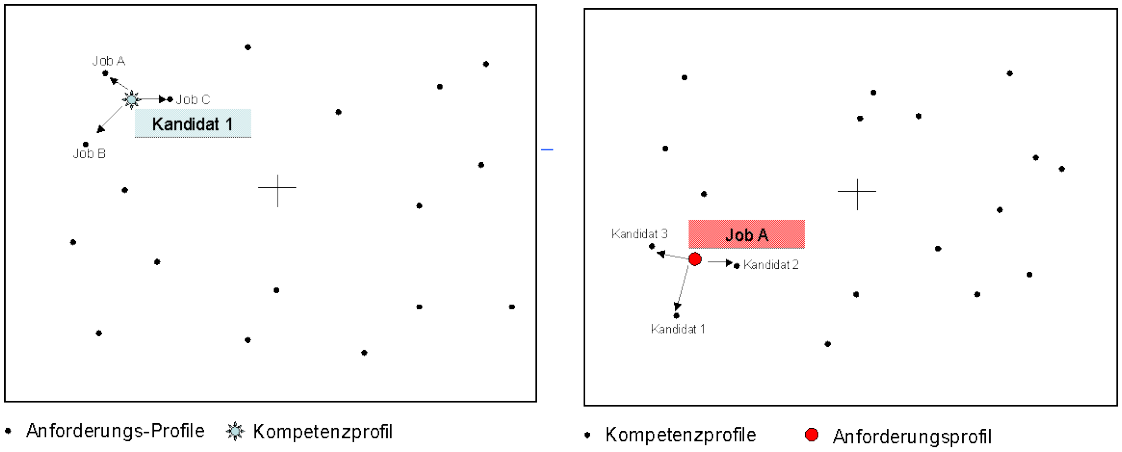



Abbildung 8: Laufbahn- und Nachfolgeplanung mittels Nonmetrischer Multidmensionaler Skalierung

Voraussetzung für die Skalierung der Daten ist die Verwendung eines gemeinsamen Kompetenzmodells sowohl für die Erstellung von Anforderungsprofilen als auch für die Messung von Kompetenzprofilen. Bei integrierten Kompetenzmanagement-Ansätzen wie sie etwa Hilb (2004) vorschlägt, ist dies auch der Fall. Zur Erstellung von Anforderungsprofilen

können in Analogie zu dem oben beschriebenen Kompetenzmessverfahren die Stelle bewertet werden, indem die Wichtigkeit der jeweiligen Kompetenz für die entsprechende Stelle in eine Rangreihe gebracht wird.

Die in dieser Studie verfolgte Grundüberlegung bei der Messung von Kompetenzen ist die bewusste Trennung von Profil und Profilhöhe. In der Messung des Profils einer Person wird lediglich über dessen relative Stärken und Schwächen in bestimmten Kompetenzdimensionen eine Aussage gemacht. Die Summe der Profile, d.h. die Summe der Ausprägungsgrade der Kompetenzdimensionen, ist demnach für jede Person dieselbe. Die absolute Profilhöhe wird dabei ausgeblendet, um Halo-Effekte

zu vermeiden.[10] Natürlich interessiert bei Personalbeurteilungsprozessen jedoch nicht nur das relative Stärken-/ Schwächenprofil einer Person, sondern auch die absolute Profilhöhe bzw. der tatsächliche Ausprägungsgrad in der jeweiligen Kompetenzdimension. So kann zum Beispiel sowohl für die Sekretärin als auch für das Geschäftsleitungsmitglied die Kommunikationsfähigkeit als absolute Stärke gelten, die Ausprägung in der jeweiligen Kompetenz ist jedoch mit grosser Wahrscheinlichkeit unterschiedlich. Nun müssten also streng genommen für jeden der 75 Kompetenzbegriffe valide Tests vorliegen, welche den Ausprägungsgrad der jeweiligen Kompetenz messen.

Da dieser Aufwand enorm hoch wäre, wollen wir ein Modell vorschlagen, welches es erlauben würde, den Testaufwand auf ein Minimum zu reduzieren: Zu diesem Zweck nutzt man die Informationen der relationalen Beziehungen zwischen den Kompetenzprofilen, welche durch die Profilerhebung gewonnen wurden. Basierend auf den Informationen aus den Beziehungen der Kompetenzen zwischen den Personen, würde es vollkommen ausreichen, wenn man drei Kompetenzen valide messen könnte, um die Höhe der restlichen Kompetenzen durch Interpolation zu schätzen. Dazu ist es von Vorteil, die gesamte Varianz einer Person zu erfassen, in dem man für eine Person sowohl den höchsten, einen mittleren als auch den tiefsten Wert innerhalb eines Kompetenzprofils valide misst. Indem man die Personen untereinander vergleicht, lässt sich anhand der drei gemessenen Kompetenzwerte ableiten, mit wie viel Prozent jemand mit seinem spezifischen Wert auf der jeweiligen Kompetenz über der Gesamtpopulation liegt. Durch diesen relativen Vergleich zwischen den Personen und der Berücksichtigung der Varianz innerhalb eines Kompetenzprofils lassen sich die Werte auf allen Kompetenzdimensionen für jede Person ermitteln.

Auf diese Weise kann die absolute Profilhöhe auf ökonomische Weise gemessen werden, indem man sich die Informationen der relativen Beziehungen zwischen den Kompetenzprofilen zu Gute macht.

Bei der Frage der Darstellung der unterschiedlichen Profilhöhe behilft man sich eines einfachen visuellen Mittels: Die Profilhöhe, d.h. die aggregierte Summe aller

[10] Die Bedeutung von Halo –Effekten bei Leistungs- und Kompetenzmessungen wurde in einer Metaanalyse von Viswesvaran, Schmidt und Ones (2005) bestätigt.

Kompetenzwerte einer Person, liesse sich in der euklidischen Karte durch die variable Grösse des Punktes des jeweiligen Kompetenzprofils darstellen.

Somit sind die Grundlagen geschaffen, welches einen ökonomischen Zugang zur Messung von Kompetenzprofilen erlaubt. Einerseits können Kompetenzprofile durch Selbstbeurteilungsdaten mittels Forced-Choice-Verfahren erhoben werden, andererseits kann zur Messung der absoluten Profilhöhe der Messaufwand auf ein paar wenige Kompetenzdimensionen reduziert werden, indem man sich auf die Informationen aus den relationalen Beziehungen der Profildaten abstützt. Nun liesse sich einwenden, dass die Messung von Kompetenzprofilen mittels Selbstbeurteilung im Allgemeinen wenig valide ist. Hier sind insbesondere die durch Tedeschi & Norman, N. (1985) gut untersuchten Selbstüberhöhungs-Effekte von Selbstbeurteilungsdaten zu erwähnen. Da die Selbstbeurteilungsdaten jedoch nicht im gängigen Fragebogenformat auf einer mehrstufigen Skala, sondern mittels Forced-Choice-Verfahren erhoben wurden, kommen diese Selbstüberhöhungseffekte nicht zum Tragen.

Um die Gültigkeit des in der Diskussion skizzierten Ansatzes des Einbezugs der Profilhöhe empirisch zu überprüfen, ist jedoch weitere Forschung wünschenswert. Als Abschluss dieser Diskussion stellen wir ein mögliches Versuchsdesign vor, welches die Validität eines solchen Ansatzes bekräftigen könnte: Mittels einer Stichprobe von mindestens 20 Personen würden in zwei Messbedingungen zwei euklidische Räume von Kompetenzprofilen erstellt und mittels Prokrustes-Transformation, einem Verfahren welches die strukturelle Abweichung zwischen zwei Karten misst, miteinander verglichen. In der ersten Messbedingung würden die Kompetenzprofile durch ein aufwändiges Assessment gemessen, welches Führungskräfte in zehn Kompetenzdimensionen evaluiert. In der zweiten Messbedingung wird dieselbe Stichprobe durch ein weniger aufwändiges Assessement gemessen, welches die Führungskräfte in lediglich drei Kompetenzdimensionen evaluiert. Zusätzlich erstellen die Führungskräfte mittels des in dieser Studie vorgestellten Forced-Choice-Verfahren Kompetenzprofile auf allen zehn Kompetenzdimensionen. Die drei Kompetenzwerte aus dem Assessment werden mit den Kompetenzprofilen in Beziehung gesetzt, indem durch Interpolation für jede Kompetenzdimension gemäss der Rangreihe aus dem Forced-Choice-Verfahren ein entsprechender Kompetenzwert berechnet wird. Im Anschluss würden, basierend auf den Daten beider Messbedingungen, euklidische Karten berechnet, welche dann miteinander verglichen

werden. Die Hypothese würde lauten: Die strukturelle Abweichung der beiden Karten ist derart gering, dass man in Zukunft die fachübergreifende Kompetenzmessung von Führungskräften deutlich vereinfachen könnte.

## 6.5 Literatur

Batram, D. (2005): The great eight competencies : A criterion-centric approach to validation *Journal of applied psychology*, 90, 6, 1185-1203.

Boyatzis, R.E. (1982). The competent manager. In A. Stewart (ed.), *Motivation and Society.* San Francisco CA.: Jossey Bass.

Boyatzis, R.E. (2006). Leadership competencies. In Ronald Burke and Cary Cooper (eds.), *Inspiring Leaders*, London: Routledge Press (Taylor & Francis Group). pp. 119-148.

Dulewicz, V. & Young, M. (2008). Similarities and Differences between Leadership and Management: High-Performance Competencies in the British Royal Navy. British Journal of Management, 19, 1, 17-32.

Erpenbeck, J. /Heyse, V./Max, H. (1999): KODE®, Berlin/Regensburg/Lakeland (Florida).

Erpenbeck, J./v. Rosenstiel, L.(Hg.) (2007): Handbuch Kompetenzmessung. Erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis, 2. Auflage, Stuttgart.

Jetter, F. & Skrotzki, R. (2005). Führungskompetenz. Düsseldorf: Metropolitan Velag 2001; erneut Regensburg: Walhalla-Fachverlag.

Gapski, H. (2001). Medienkompetenz. Eine Bestandsaufnahme und Vorüberlegungen zu einem systemtheoretischen Rahmenkonzept, Verlag für Sozialwissenschaten; 1. Aufl.

Hilb, M. (2006). Integriertes Personalmanagement. Luchterhand Verlag GmbH; Auflage: 15., aktualis. Auflage.

Kalpaka, A. (1998): „Interkulturelle Kompetenz. Kompetentes (sozial)pädagogisches Handeln in der Einwanderungsgesellschaft“, in: *IZA, Zeitschrift für Migration und Soziale Arbeit*, Heft 3-4, S.77-79, Frankfurt/Main.

Läge, D., Daub, S., Bosia, L., Jäger, C., & Ryf, S. (2005). Die Behandlung ausreisser behafteter Datensätze in der Nonmetrischen Multidimensionalen Skalierung - Relevanz, Problemanalyse und Lösungsvorschlag (Forschungsberichte aus der Angewandten Kognitionspsychologie Nr. 21).Universität Zürich, Psycholosches Institut.

North, K. (1999). Wissensbasierte Unternehmensführung: Wertschöpfung durch Wissen. 2. Aufl. Wiesbaden: Gabler.

Probst, G. et al. (1999). Wissen managen: Wie Unternehmen ihre wertvollste Resource optimal nutzen. Wiesbaden: Gabler.

Probst, G. et al. (2000). Kompetenz Management. Wie Individuen und Organisationen Kompetenz entwickeln. 1. Aufl. Wiesbaden: Gabler.

Schippmann, J.S. (1999). The practice of competency modelling. *Personnel Psychology*, 53, 703-740.

Schuler, H. & Funke, U. (1995). Diagnose beruflicher Eignung und Leistung. In H. Schuler (Hrsg.), *Lehrbuch Organisationspsychologie* (2. Aufl., S. 235-284). Bern: Huber.

Spencer, L.M., Jr. & Spencer, S.M. (1993). Competence At Work - Models For Superior Performance. New York: Wiley.

Spencer, L.M., Jr. & Spencer, S.M. (2008). Competence At Work - Models For Superior Performance. New York: Wiley.

Tedeschi, J. T., & Norman, N. (1985). Social power, self-presentation, and the self. In B. R. Schlenker, (Ed.), *The self and social life* (pp. 293–322). New York: McGraw-Hill.

Thom, N. & Zaugg, R.J.. (2001). Excellence durch Personal- und Organisationskompetenz. Haupt Verlag.

Viswesvaran, C., Schmidt, F.L. & Ones, D.S. (2005). Is There a General Factor in Ratings of Job Performance? A Meta-Analytic Framework for Disentangling Substantive and Error Influences. *Journal of Applied Psychology*, 90, 1,108-131.

Wildmann, L. (2001): Der Kompetenzmensch. Lernen – und das ein Leben lang, Sternenfels.

Wunderer, R. (2007). Führung und Zusammenarbeit - eine unternehmerische Führungslehre. 7. überarbeitete Auflage. Köln : Luchterhand.

# 7 Kompetenzmodellierung mittels Nonmetrischer Multidimensionaler Skalierung

## 7.1 Einleitung

„Testing for Competence Rather Than for „Intelligence“ war der Titel eines Artikels von McClelland (1973), der den Grundstein für die beachtliche „Competency-Bewegung der letzten drei Jahrzehnte gelegt hat. Die Schlagworte „Kompetenz“ bzw. „Kompetenzmodelle“ – insbesondere in Bezug auf einen berufsorientierten Handlungskontext, hat in den vergangenen Jahren die (wirtschafts-) pädagogische sowie arbeits- und organisationspsychologische Debatte entscheidend beeinflusst und mitgeprägt. Insbesondere auf die Unterscheidung zwischen traditioneller und vergangenheitsorientierter Anforderungsanalyse und zukunftsgerichteter Kompetenzmodellierung wurde von Sarges (2001) ein massgebender Artikel publiziert, welcher die Bedeutung von Kompetenzmodellen für die Praxis herausstreicht. Laut Sarges entwickelten sich aus der angelsächsischen Tradition der Competency–Bewegung zwei gewichtige Vorteile gegenüber der klassischen Anforderungsanalyse:

1. Mit der Formulierung von Kompetenzen erlaubt man sich eine grössere Unbefangenheit in Richtung auf die Alltagssprache.
2. Man kann die Zukunft explizit mit einbeziehen, die Competencies auf die Unternehmensstrategie beziehen und teilweise sogar einen breiteren Bezugsrahmen für viele wichtige HR-Aktivitäten wie Personalselektion und –Entwicklung erhalten.

So verwundert es nicht, dass sich eine wahre Competency–Euphorie entwickelt hat, welche dazu führte, dass in jedem grösseren Unternehmen ein unternehmensweit angewendetes Kompetenzmodell zu Grunde gelegt wird, welches dazu dienen soll, die fachübergreifenden Kompetenzen der Mitarbeitenden bzw. Führungskräfte zu messen, um daraus folgende Personalmassnahmen abzuleiten.

Weltweit gibt es eine ganze Menge an sogenannten „Competency Models“, spezifische und generelle, mehr aus der akademischen Welt kommende und solche aus der Praxis. Die praxis-basierten Modelle sind in der Regel weiter entwickelt und detaillierter in Richtung auf Instrumentierung, Verhaltensanker und assoziierte

Entwicklungsinstrumente. Die akademischen Modelle dagegen versuchen, mittels gängigen dimensionsreduzierenden Verfahren wie Faktoren- oder Clusteranalyse eine kleinere Anzahl von generellen Dimensionen zu finden, die ein umfassendes und dennoch sparsames Instrumentarium für die Domäne berufsrelevanter Anforderungsmerkmale darstellen. Laut Kurz & Bartram, (2001) werden nun immer dringlicher Kompetenzmodelle benötigt, welche die Sparsamkeit und Struktur der akademischen Modelle mit der Brauchbarkeit und Praktikabilität der angewandten Modelle aus der Praxis kombinieren. Neben der Schwierigkeit Kompetenzmodelle zu konstruieren, die aus Praktikabilitätsgründen nicht zu umfassend und schwerfällig, jedoch zur differenzierten Beschreibung von Persönlichkeitstypen im Management eine hinreichend grosse Anzahl an Kompetenzen bieten, liegt die zweite Herausforderung in der Messung der jeweiligen Kompetenzen.

Kompetenzeinschätzungen im Rahmen von Personalbeurteilungen werden meistens im gängigen Fragebogenformat durchgeführt. Mithilfe von Ratingskalen werden in Bezug auf die Führungskompetenzen die Ausprägungsgrade „sehr stark ausgeprägt“ bis „gar nicht ausgeprägt“ abgefragt. Die Gefahr solcher Fragebogenformate sind einerseits die kognitive Überlastung durch gleichzeitige Messung von Profil und absoluter Profilhöhe (Vgl. Forschungsbericht I, II und III der vorliegenden Arbeit) und andererseits die Verzerrung der Werte durch die Anwendung eines idiosynkratischen Beurteilungsmassstabes der verschiedenen Beurteiler. Rangordnungsskalen erfordern eine Einreihung von Items. Die Befragten werden hierbei gebeten eine Rangordnung zwischen verschiedenen Antwortalternativen zu erstellen. Der Vorteil von Rangordnungsskalen ist die relative Einschätzung eines Items zu anderen. Damit können gerade im Hinblick auf die Profilbildung differenziertere Kompetenzurteile abgegeben werden, als mittels Ratingskalen auf einem Merkmalskontinuum.

Die vorliegende Studie umfasst demnach zwei Ziele: Zum einen gilt es zu zeigen, dass ipsative Messung in Form eines Forced-Choice-Verfahrens als eine echte Alternative zu den gängigen Ratingskalen normativer Messung betrachtet werden kann. Zum anderen soll gezeigt werden, wie sich eine Managerstichprobe anhand einer bestimmten Anzahl fachübergreifender Kompetenzen hinreichend differenziert und stabil beschreiben lässt, unabhängig davon welche Terminologie man in den einzelnen Kompetenzmodellen anwendet.

## 7.2 Methodik

In einem ersten Teil stellen wir das dieser Untersuchung zu Grunde liegende Kompetenzmodell vor und erklären in einem zweiten Teil die Messung von Kompetenzprofilen mittels eines onlinebasierten Forced-Choice-Ansatzes. Dabei werden die Hypothesen aus den Fragestellungen abgeleitet sowie die angewendeten Methoden zur Überprüfung der Hypothesen vorgestellt.

*Theoriegeleitetes Kompetenzmodell*

Zur Entwicklung des theoriegeleiteten Kompetenzmodells wurde einerseits die in der Praxis angewendeten Kompetenzmodelle[11] auf Gemeinsamkeiten evaluiert sowie die 30 jährige Kompetenzforschung (Vgl. Spencer (2008) Boyatzis (2006), Schippmann (2000); Dulewicz & Young (2008)) berücksichtigt.

Dabei hat sich gezeigt, dass sich die Inhalte der Kompetenzmodelle sowohl aus theoretischer als auch aus praktischer Sicht weitgehend wiederholen, die Struktur bzw. Clusterung der Kompetenzen sich jedoch ändert. So findet man z.B. bei Boyatzis die drei Cluster Goal and Action Management Cluster, People Management Cluster und Analytic Reasoning Cluster während man bei Wunderer (2007) die Cluster Gestaltungskompetenz, Sozialkompetenz und Umsetzungskompetenz findet, während wiederum andere Wissenschaftler die Begriffswelt der fachübergreifenden Kompetenzen in vier (Erpenbeck & Von Rosenstiel 2007) bis acht kategoriale Cluster einteilen (Vgl. Bartram, 2001).

Die meisten grösseren Beratungsfirmen bieten inzwischen hierarchische Modelle an: mit einer kleinen Zahl breiter Faktoren bzw. Cluster oben und einer grossen Menge von Komponenten oder Elementen unten. Abbildung 3 zeigt dies synoptisch für die Beratungsfirmen DDI, PDI und SHL:

[11] Bei der Analyse wurden branchenfremde Modelle von Siemens, ABB, IBM, KMPG, Daimler Chrysler, Credit Suisse sowie das Schweizerische Militär berücksichtigt, während bei der branchenspezifischen Analyse die Kompetenzmodelle der CKW, Axpo Informatik AG, NOK, EGL und Stadtwerke Münster einbezogen wurden.

| **DDI** | **PDI** | **SHL** |
|---|---|---|
| 6 Clusters<br>e.g. Expressing Individual Potential | 8 Factors<br>e.g. Thinking Skills | 8 Factors<br>e.g. Interacting & Presenting |
| Ca. 50 Dimensionen/ Competencies<br>e.g. Adaptability | Ca. 25 Competencies<br>e.g. Analytical Thinking | 20 Dimensions<br>e.g. Persuading & Influencing |
| Key Actions<br>e.g. Approaches Change positively | Behaviors<br>e.g. Systematically gathers relevant information | Competency Components<br>e.g. Gaining agreement |

Abbildung 1: Synoptischer Vergleich der hierarchischen Competency-Kataloge der Firmen DDI, PDI und SHL (Quelle, Sarges, 2001, S. 10)

Basierend auf der Analyse der gängigen Kompetenzmodelle in der Praxis und den verschiedenen Kompetenzmodellen aus wissenschaftlicher Forschung wurde für die empirische Untersuchung innerhalb des für diese Studie herangezogenen Energieunternehmens ein Kompetenzmodell postuliert, welches ebenfalls hierarchisch aufgebaut ist und alle Kompetenzklassen nach Erpenbeck und Von Rosenstiel (2007) abdeckt:

- Personale Kompetenzen
- Aktivitäts- und umsetzungsbezogene Kompetenzen
- Fachlich-methodische Kompetenzen
- Sozial-kommunikative Kompetenzen

Die Kategorisierung der Kompetenzbegriffe in 4 Kompetenzfelder, 15 Kompetenz-Dimensionen und 60 Kompetenzfacetten ist rein hermeneutisch und basiert auf der Zuordnung von Personalfachleuten und Linienmanagern eines Energieunternehmens. Die Zuteilung hat somit rein praktischen Nutzen und basiert nicht auf einer faktorenanalytischen Kategorisierung.

| Fokus - Erfolgsstrategien entwickeln | | Leistungsverhalten - Initative ergreifen und in hoher Qualität umsetzen | |
|---|---|---|---|
| **Unternehmerisches Denken und Handeln** | Marktorientierung<br>Chancen-/Risikobewusstsein<br>vernetztes Denken<br>Kosten-/Nutzen Denken | **Leistungsmotivation** | Ambition<br>Initiative<br>Verantwortungsübernahme<br>Engagement |
| **Problemlösungsfähigkeit** | Analytische Fähigkeiten<br>system.-methodisch. Vorgehen<br>Schlussfolgerndes Denken<br>Urteilsvermögen | **Selbstmanagement** | Umgang mit eigenen Ressourcen<br>Selbstreflexionsfähigkeit<br>Belastbarkeit<br>Überblick bewahren |
| **Planungs-und Organisationsfähigkeit** | Projektmanagement<br>Prozessdenken<br>Finanz- und Ressourcenmangt<br>Prioritäten setzen | **Umsetzungsorientierung** | Entscheidungsfähigkeit<br>Ergebnisorientierung<br>Durchsetzungsvermögen<br>Beharrlichkeit |
| **Offenheit für Neues** | Antizipationsfähigkeit<br>Flexibilität<br>Lernfähigkeit<br>Innovationskraft | **Qualitätsbewusstsein** | Verbesserungsstreben<br>Verlässlichkeit<br>Genauigkeit<br>Sicherheitsbewusstsein |
| **Kundenorientierung** | Kundenbedürfnisse erkennen<br>Beratungsfähigkeit<br>Kundenfreundlichkeit<br>Netzwerke pflegen | | |
| **Soziale Kompetenz - Partner gewinnen** | | **Leadership - Team führen** | |
| **Kommunikationsfähigkeit** | Information & Rückmeldung<br>Verhandlungsfähigkeit<br>Gesprächsführung<br>Präsentationsfähigkeit | **Andere Inspirieren** | Sinn vermitteln<br>Veränderungsbereitschaft<br>Integrität<br>Begeisterungsfähigkeit |
| **Zusammenarbeit / Teamfähigkeit** | Teamorientierung<br>Hilfsbereitschaft<br>Integrationsfähigkeit<br>Kooperationsfähigkeit | **Führungsverhalten** | Zielorientiertes Führen<br>Fähigkeit zur Delegation<br>Teamleistung fördern<br>Zielerreichung überprüfen |
| **Konfliktlösungsfähigkeit** | Zwischenmenschliches Feingespür<br>Perspektivenübernahme<br>Konfliktfreudigkeit<br>Umgang mit Kritik | **Mitarbeiterförderung** | Vertrauen schaffen<br>Feedback geben<br>Leistung einfordern<br>Andere fördern |

Abbildung 2: hermeneutisches Kompetenzmodell als Datengrundlage der empirischen Studie

Das Modell wurde dann post hoc einer empirischen Untersuchung unterzogen, welche Aufschluss über die inhaltliche Struktur der verschiedenen Kompetenzbegriffe sowie Aufschluss über die Anzahl Dimensionen eines ökonomischen und gleichzeitig differenzierenden Kompetenzmodells geben soll. Bevor auf die empirische Untersuchung eingegangen wird, soll auf die Methodik der verschiedenen Kompetenzmessverfahren bzw. Persönlichkeitstests eingegangen werden. Dabei werden auch die in dieser Studie angewendete Methodik erläutert und die interessierenden Fragestellungen formuliert.

*Verfahren zur Messung von Kompetenzen*

In der Psychologie werden seit jeher verschiedene Verfahren zur Kompetenzmessung angewendet. Erpenbeck und Rosenstiel (2007) stellen in ihrem Handbuch der Kompetenzmessung 50 verschiedene Verfahren vor. Auch die Persönlichkeitspsychologie und die Forschung zur Emotionalen Intelligenz (Boyatzis, Goleman & Ree, 2007), haben zahlreiche psychometrische Verfahren zur Messung berufsrelevanter Kompetenzen entwickelt. Beispiele gut validierter Verfahren sind das von Hossiep & Paschen (2003) entwickelte Bochumer Inventar zur berufsbezogenen

Persönlichkeitsbeschreibung (BIP), der von Saville & Holdsworth (2003) revidierte Occupational personality questionnaire (OPQ32) oder das von Goleman durch langjährige Forschung der emotionalen Intelligenz entwickelte Emotional Competency Inventory (ECI). Ausser dem OPQ-Verfahren, welches auf einem Forced-Choice-Antwortformat basiert, werden die meisten Persönlichkeitstests oder Kompetenzmessverfahren auf Intervallskalen erhoben. Die Rohwerte werden dann in Bezug zu Normstichproben gesetzt, um damit Aussagen über ein bestimmtes berufsrelevantes „Persönlichkeitsprofil" machen zu können. In diesen Verfahren wird durch die Messung demnach nicht nur das Profil, sondern auch die absolute Profilhöhe gemessen. Dadurch ergeben sich erhebliche Messfehler aufgrund eines idiosynkratisch angewendeten Massstabs. Tendenz zur Milde, Strenge und Mitte sind hier, neben den von Tedeschi und Norman (1985) bekannten Effekten der positiven Selbstdarstellung, die gängigen Messfehler, welche aufgrund des Fragebogenformats zu Stande kommen.

In der vorliegenden Untersuchung wurde ein Verfahren entwickelt, welches lediglich das Profil und nicht die Profilhöhe berufsrelevanter Kompetenzen abfragt. Das Verfahren sieht vor, aufgrund eines Forced-Choice-Verfahrens Kompetenzprofile zu bilden, welche die Stärken einer Person abbildet und diese dann mittels Korrelationskoeffizienten zu anderen Personen in Beziehung setzt.

Dabei können die mittels des Forced-Choice-Verfahrens erhobenen Daten sowohl auf Personenebene als auch auf Kompetenz- bzw. Itemebene mittels des strukturentdeckenden und visualisierenden Verfahren der Nonmetrischen Multidimensionalen Skalierung (NMDS) anhand eines robusten Algorithmus[12] ausgewertet werden. Die Struktur der Daten lässt sich in einer kognitiven Karte darstellen, wobei die Distanzen zwischen den Objekten deren Relationen widerspiegeln. Somit werden ähnliche Objekte, die untereinander hoch kovariieren,

[12] Läge, D., Daub, S., Bosia, L., Jäger, C., & Ryf, S. (2005). Die Behandlung ausreißerbehafteter Datensätze in der Nonmetrischen Multidimensionalen Skalierung - Relevanz, Problemanalyse und Lösungsvorschlag (Forschungsberichte aus der Angewandten Kognitionspsychologie Nr. 21). Universität Zürich, Psychologisches Institut.

nahe beieinander abgebildet, während tief kovariierende Objekte weit voneinander zu liegen kommen. Der Vorteil in der Darstellung der NMDS liegt darin, dass durch die Visualisierung der Daten die Nachbarschaftsbeziehungen zwischen den verschiedenen Objekten auf einen Blick ersichtlich werden. Anders als bei den gängigen dimensionsreduzierenden Analyseverfahren wie z.B. bei der Faktorenanalyse wird bei der NMDS die gesamte Varianz berücksichtigt, ohne dass Informationen systematisch verloren gehen.

In einem ersten Schritt (Vgl. Kapitel 5.1) werden die Personen bzw. die Kompetenzprofile von Managern zueinander in Beziehung gesetzt. Mittels NMDS lassen sich Personenräume erstellen, welche die Ähnlichkeit bzw. Unähnlichkeit von Kompetenzprofilen in einem zweidimensionalen euklidischen Raum abbildet. Personen mit ähnlichen Kompetenzprofilen kommen nahe beieinander zu liegen, während unähnliche Kompetenzprofile weit auseinander abgebildet werden. Es interessiert die Frage, in wie fern solche Personenräume stabil sind bzw. von der Auswahl der gemessenen Kompetenzen abhängen.

Wir stellen an dieser Stelle die Hypothese auf, dass sich durch das in dieser Erhebung angewendete Forced-Choice-Verfahren relativ stabile Kompetenzprofile von Personen erstellen lassen, die unabhängig von der Auswahl fachübergreifender Kompetenzen ist.

Zur Überprüfung dieser Hypothese wurden nach der split-half Methode zwei Personenräume berechnet, welche jeweils auf der Hälfte der Kompetenzfacetten erstellt wurden. Dabei wurde für die eine Hälfte der Personen jeweils die 1. und die 3. Kompetenzfacette pro Kompetenz gewählt, während für die andere Hälfte jeweils die 2. und die 4. Kompetenzfacette in die Stichprobe genommen wurde.

In einem zweiten Schritt (Vgl. Kapitel 5.2) werden die Daten auf Itemebene, d.h. auf Ebene der Kompetenzbegriffe, präsentiert. Es geht um die Frage, in wie fern die Kompetenzen inhaltlich zusammenhängen bzw. um die Frage, ob sich einzelne Kompetenzcluster herausbilden. Wir gehen davon aus, dass sich in der euklidischen Karte einzelne Kompetenzcluster zeigen, die inhaltlich Ähnliches messen.

In einem dritten Schritt (Vgl. Kap. 5.3) werden dann die Kompetenzdimensionen verschiedener Kompetenzmodelle mit dem Personenraum in Beziehung gesetzt. Mittels Property Fitting, einem multiplen Regressionsverfahren, werden die Kompetenzdimensionen in den Personenraum gelegt. Dabei liegt der Schwerpunkt des Forschungsinteresses in der Anwendung verschiedener Kompetenzmodelle in

einem Personenraum bzw. in der Frage in wie weit verschiedene Kompetenzmodelle austauschbar sind.

*Empirische Erhebung*

Anhand des theoretisch hergeleiteten Kompetenzmodells wurde innerhalb mehreren Unternehmensbereichen des erwähnten Energieunternehmens eine empirische Erhebung durchgeführt. Dabei wurden Führungskräfte aus dem mittleren und oberen Managements aus unterschiedlichen Fachbereichen mittels eines onlinebasierten Kompetenzerhebungsverfahren befragt. Das Verfahren basiert, wie im methodischen Teil erwähnt, auf einem Forced-Choice-Verfahren, welches die Rangreihenbildung von Kompetenzen vorsieht und somit eher als Forced-Ordering-Verfahren bezeichnet werden sollte.

In einem ersten Schritt wurden die Versuchspersonen gebeten, die 15 Kompetenzen in drei Kategorien zu teilen, je nach Selbsteinschätzung des Ausprägungsgrades, von „weniger stark ausgeprägt“ bis „sehr stark ausgeprägt“.

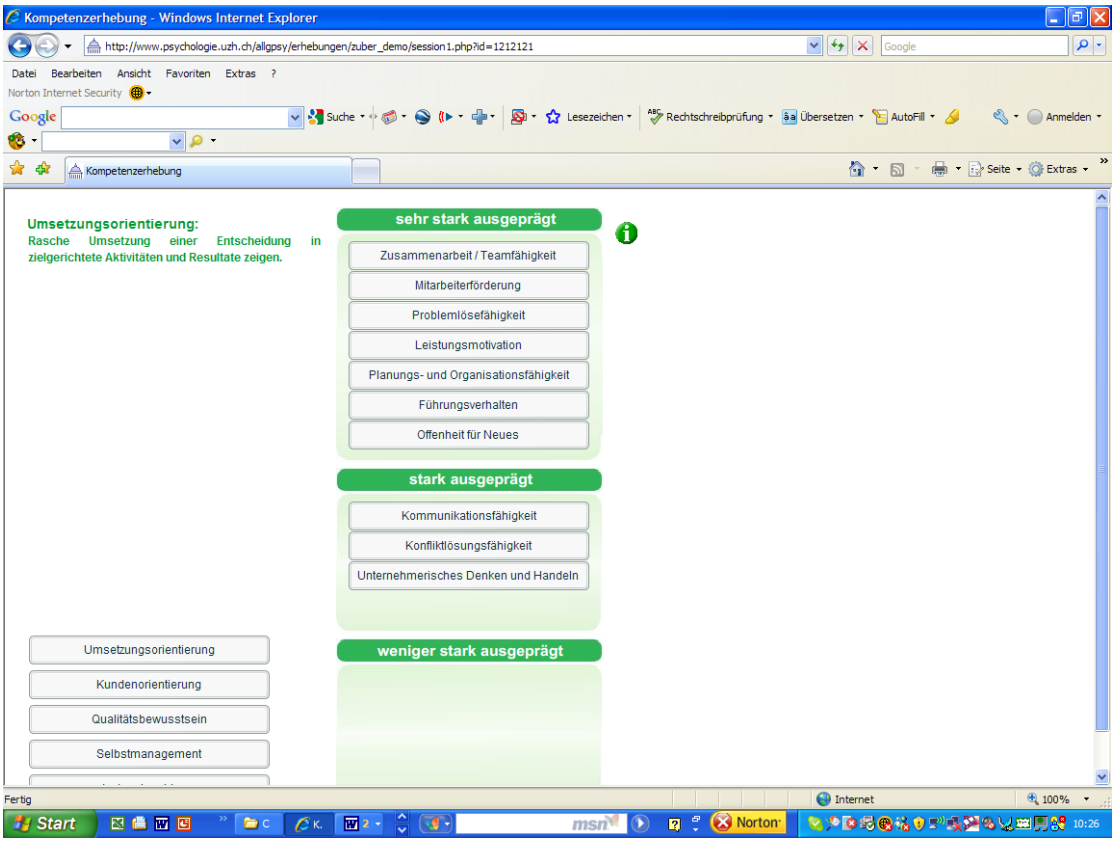



Abbildung 3: Kompetenzprofilbildung durch onlinebasiertes Forced-Choice-Tool, 1. Schritt

In einem zweiten Schritt wurden die Versuchspersonen gebeten, die Kompetenzen pro Kategorie in eine Rangreihe zu bringen. Durch dieses Verfahren wurde pro Versuchsperson ein auf Rangplatz basierendes Kompetenzprofil erstellt. Somit wurde die Datenbasis für Rangkorrelationen zwischen Kompetenzprofilen geschaffen.

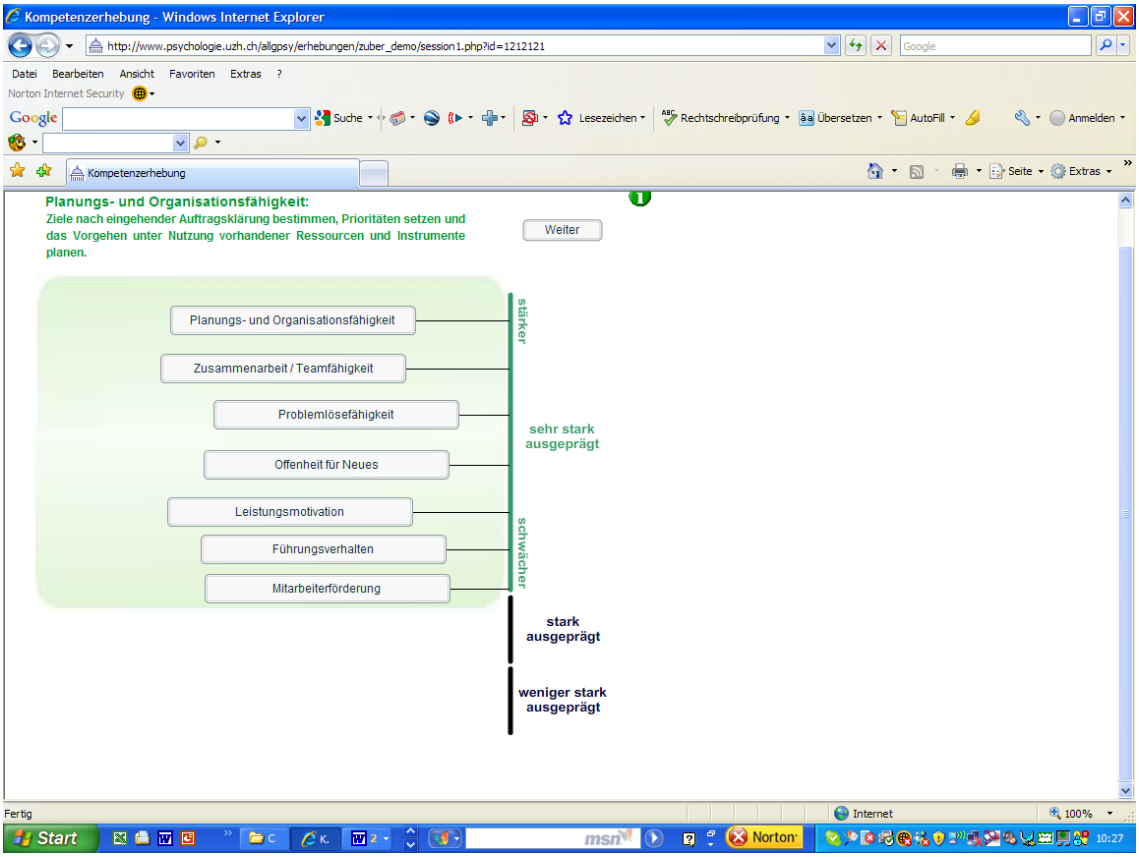



Abbildung 4: Kompetenzprofilbildung durch onlinebasiertes Forced-Choice-Tool, 2. Schritt

Basierend auf Selbsteinschätzungen zu zwei Erhebungszeitpunkten im Abstand von 2 Wochen wurden insgesamt 41 Kompetenzprofile erhoben, wobei 28 Versuchspersonen die Messbedingungen[13] erfüllten und somit in die Auswertung flossen. Alle 75 Kompetenzbegriffe (15 Kompetenzen +60 Kompetenzfacetten) wurden in der Selbsteinschätzung gemäss dem oben beschriebenen Verfahren abgefragt. Neben den 75 Kompetenzbegriffen des Kompetenzmodells wurden 15 weitere Kompetenzbegriffe gewählt, welche nicht a priori den Kategorien zuzuordnen waren. Z.B. Mobilität, Sachlichkeit oder Selbstbewusstsein. Diese Items wurden als

[13] Als Messbedingung wurden folgende Kriterien herangezogen: Teilnahme an beiden Erhebungszeitpunkten, Dauer der Durchführung = mind. > 15 Minuten

Kontrollvariablen eingeführt, um die Stabilität des Personenraums nicht in Abhängigkeit eines allzu kategorialen Kompetenzkatalogs zu messen.

## 7.3 Ergebnisse

Die Ergebnisse werden zur Beantwortung der eingangs erwähnten Fragestellungen in drei Schritten vorgestellt:

1. Ergebnisse auf Personenebene
2. Ergebnisse auf Ebene der Kompetenzen
3. Ergebnisse aus der Kombination von Personen- und Kompetenzebene

*Ergebnisse auf Personenebene*

Vergleicht man die Ähnlichkeit der Kompetenzprofile zwischen den einzelnen Versuchspersonen, korreliert man demnach die Daten auf Personenebene, so ergibt sich die unten abgebildete Karte.

Auf den ersten Blick zeigt sich eine relative Gleichverteilung über den Raum, ohne ausgeprägte Clusterung.

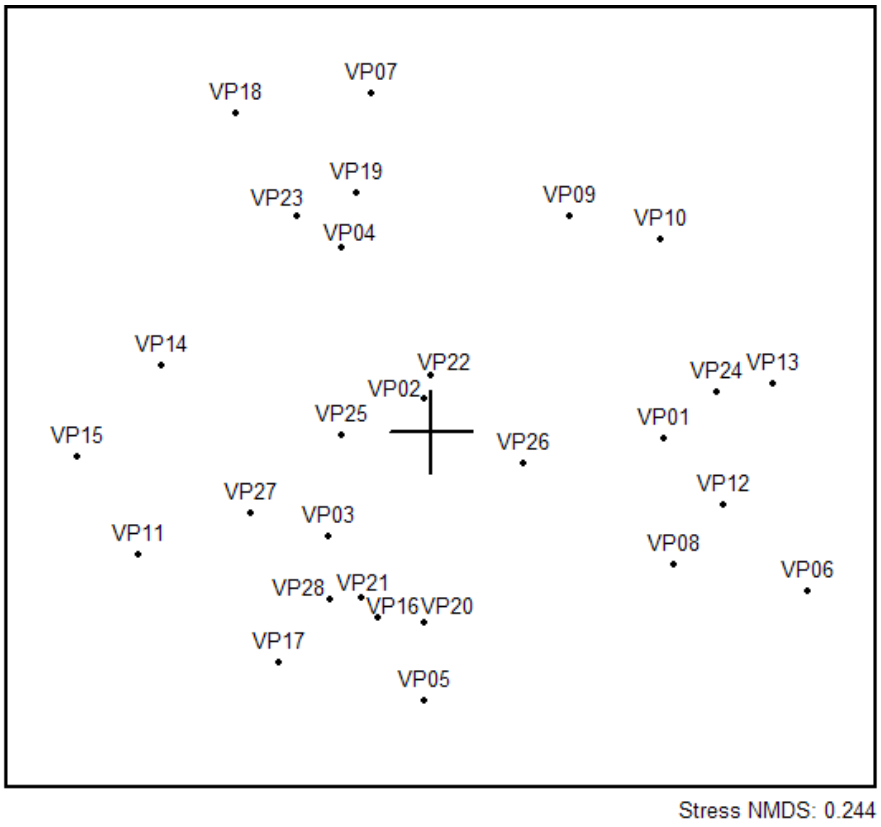



Abbildung 5: skalierte Kompetenzprofile auf Personenebene

Um die NMDS Karte sinnvoll interpretieren zu können, werden die Cluster mit den Funktionen der Versuchspersonen in Beziehung gesetzt, um die inhaltliche Plausibilität der Karte zu begründen.

Zu diesem Zweck wurde die Personenkarte in „Funktionscluster“ eingefärbt. In der unten dargestellten Abbildung lässt sich unschwer erkennen, dass sich Personen mit inhaltlich ähnlichen Funktionen in der Karte gruppieren. Oben in der Karte liegen die stark analytisch-technisch geprägten Funktionen, rechts eher die sozial ausgerichteten Funktionen wie Human Resources, Kommunikation und Marketing. Im unteren Teil der Karte kommen eher Personen zu liegen, die strategische Positionen inne haben und eher in eher höheren Hierarchieebenen zu finden sind.

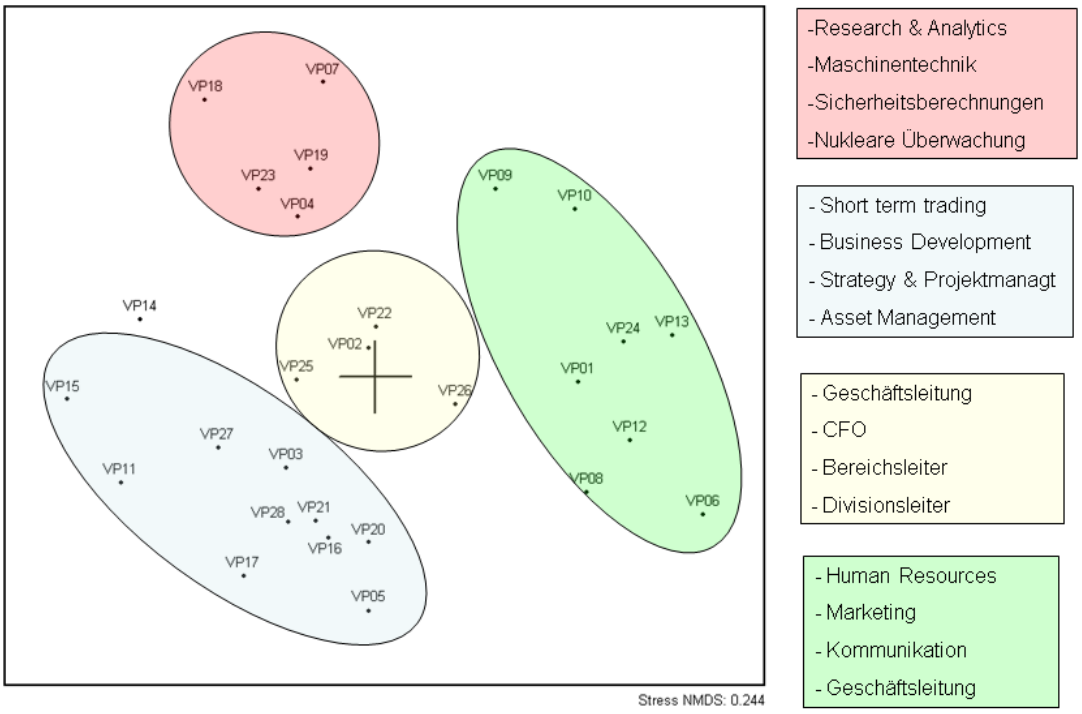



Abbildung 6: Eingefärbte Personencluster in Abhängigkeit ihrer Funktion

Interessant wird nun die Beantwortung nach der Stabilität dieses Personenraumes in Bezug auf die zu Grunde liegenden Kompetenzfacetten. Wie im methodischen Teil erwähnt, wurde durch split-half der Daten (Stichprobe 1: erste und dritte Kompetenzfacette pro Kompetenzdimension, Stichprobe 2: zweite und vierte Kompetenzfacette pro Kompetenzdimension) zwei Personenkarten berechnet. Zur Überprüfung der strukturellen Ähnlichkeit der beiden Personenkarten wurde eine Prokrustes-Transformation gerechnet. Anhand der Prokrustes-Transformation ist es möglich, die strukturelle Ähnlichkeit zweier Karten miteinander zu vergleichen.

Durch Schieben, Drehen, Spiegeln und Skalieren werden dabei die Unterschiede zwischen den beiden Karten minimiert, ohne die jeweiligen Strukturen zu verändern. Die beiden Ausgangskarten werden zusammen in einer Karte dargestellt und die korrespondierenden Punkte mit einer Line verbunden. Zudem wird mit dem "AvgLoss" ein quantitatives Mass für die Abweichung angegeben.

Abbildungen 7 und 8 zeigen die beiden Personenkarten der split-half Daten nach Prokrustes-Transformation, jeweils verglichen mit der ursprünglichen Karte bestehend aus allen Kompetenzfacetten.

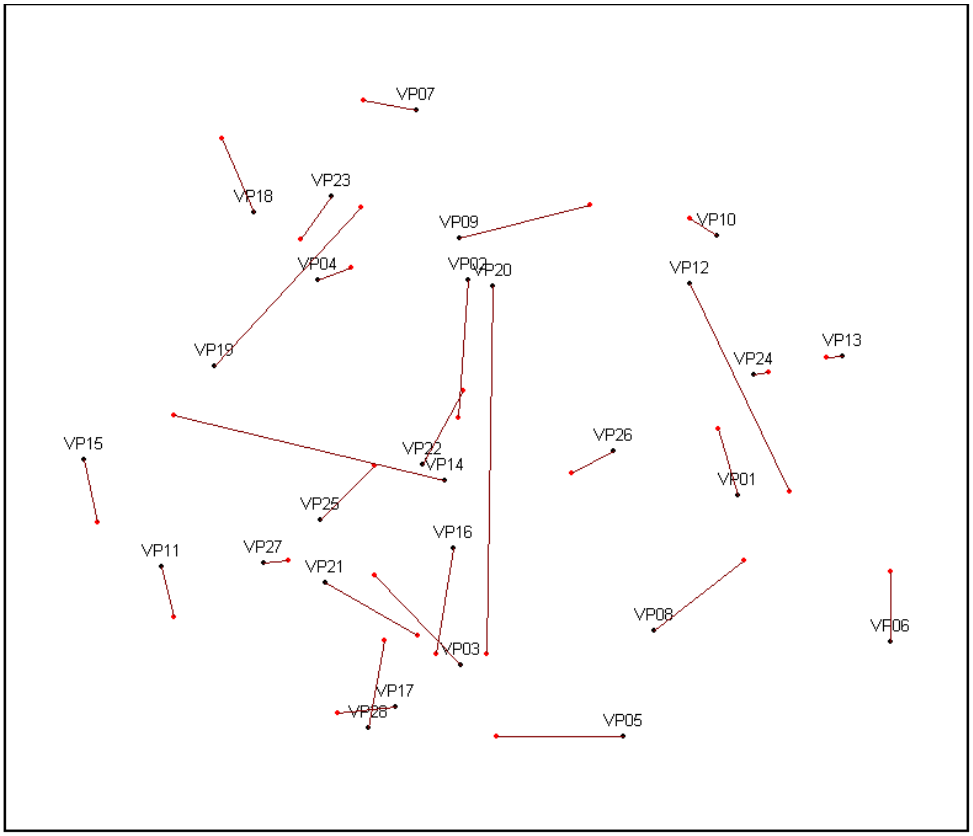



Abbildung 7: Prokrustes-Transformation der Personenkarte „alle Items“ mit Personenkarte „split-half 1. und 3. Facette“

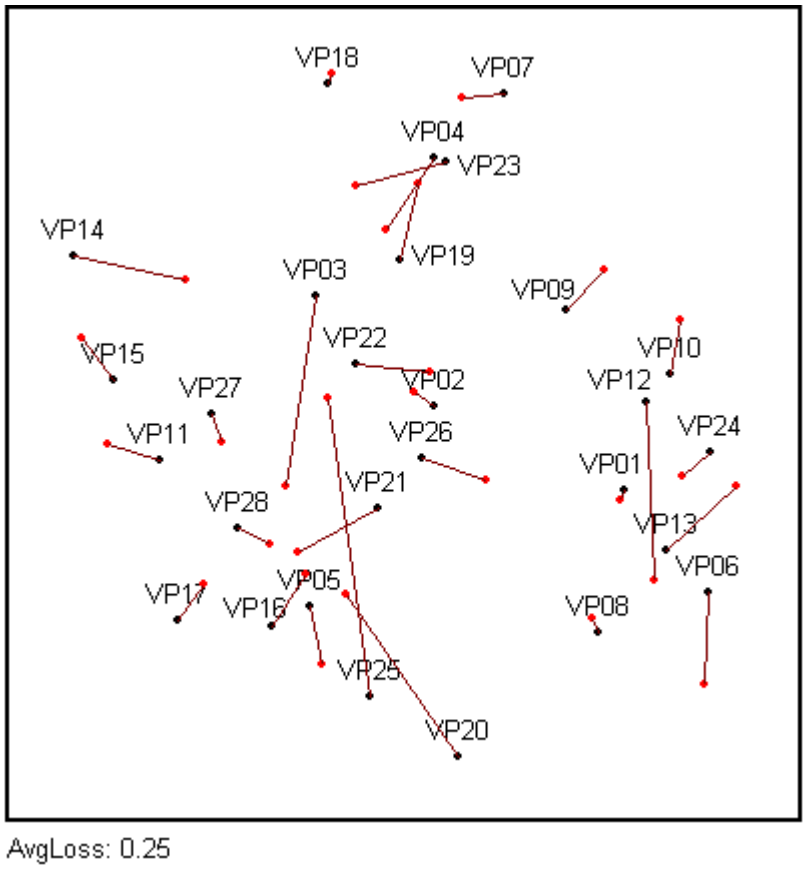



Abbildung 8: Prokrustes-Transformation der Personenkarte „alle Items“ mit Personenkarte „split-half 2. und 4. Facette“

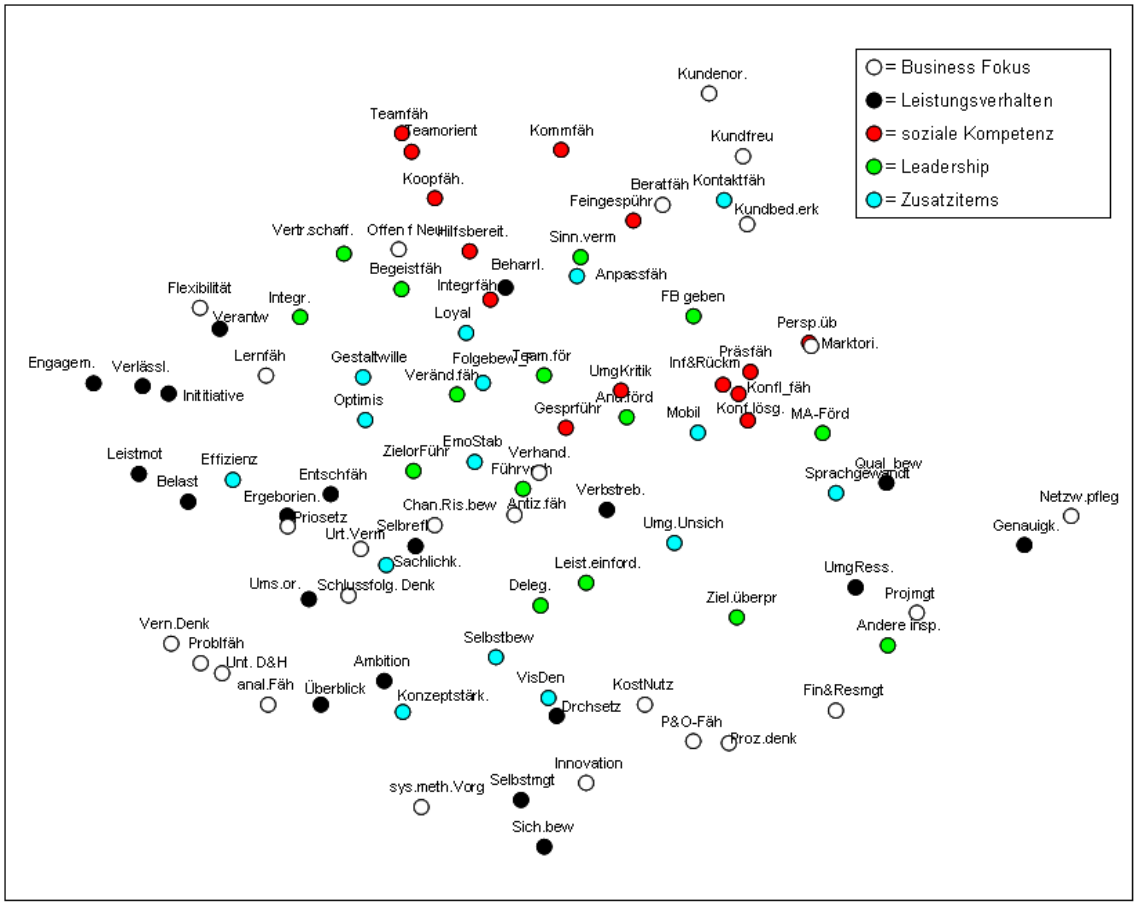
Die Ergebnisse der Prokrustes-Transformationen weisen auf eine relativ hohe Stabilität des Personenraumes hin. Die Average Losses unter .3 können als relativ tief bezeichnet werden, und die Abweichung ist hauptsächlich auf ein paar wenige „Springer“ (VP 20, VP 25, VP14 und VP 12) zurückzuführen. Die Messung des Personenraumes mittels des hier angewendeten Forced-Choice-Verfahrens ist somit relativ stabil und nicht von einzelnen Kompetenzfacetten abhängig. Welche Implikationen dieses Ergebnis auf die Messung von Kompetenzprofilen hat, wird in der Diskussion weiter erläutert.

*Ergebnisse auf der Ebene der Kompetenzen*

Nachdem auf Ebene der Personen gezeigt werden konnte, dass sich basierend auf dem Forced-Choice-Verfahren ein relativ stabiler Personenraum messen lässt, interessiert die Frage, wie die einzelnen Kompetenzen und Kompetenzfacetten, in der Folge als Kompetenzitems bezeichnet, zusammenhängen. Für diese Fragestellung wurden wiederum mittels NMDS euklidische Karten berechnet, welche die Ähnlichkeiten bzw. Unähnlichkeiten zwischen den einzelnen Kompetenzitems abbilden.

Abbildung 9 zeigt die euklidische Karte der 89 Kompetenzitems. Die Kompetenzitems sind in der a priori hermeneutisch angenommenen Kompetenzstruktur der 4 Kompetenzfelder Business Fokus, Leistungsverhalten, Soziale Kompetenz und Leadership eingefärbt. Zudem ist die Verteilung der Zusatzitems (blaue Punkte) ebenfalls in der Karte ersichtlich.

Auf den ersten Blick fällt auf, dass sich die empirisch erhobene Kompetenzstruktur nicht vollständig in den angenommenen Kompetenzkategorien lesen lässt. Es ergibt sich vielmehr eine Gleichverteilung ohne klar abgrenzbare Kompetenzcluster. Dennoch lassen sich bei genauerem Betrachten gewisse inhaltliche Schwerpunkte erkennen. Oben in der Karte liegen eher die Kompetenzitems, welche das Konstrukt der sozialen Kompetenz abdecken. Links in der Karte liegen Kompetenzitems, die eher motivationalen bzw. handlungsorientierten Charakter haben, während in der unteren Hälfte der Karte eher kognitiv gefärbte Kompetenzitems zu liegen kommen, die in der Begriffswelt der gängigen Kompetenzmodelle in der Praxis zu den „methodischen und unternehmerischen Kompetenzen“ zählen.



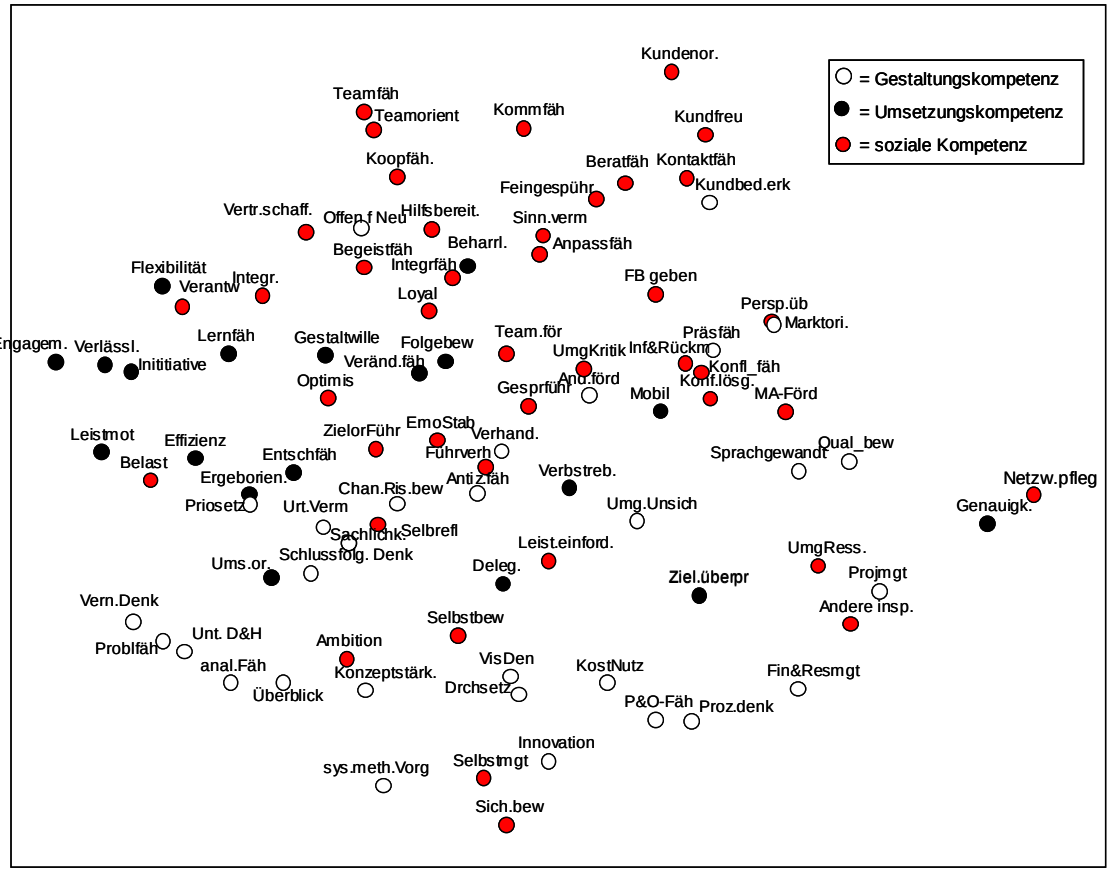

Stress NMDS: 0.224

Abbildung 9: „Kompetenzlandschaft“ bestehend aus Kompetenzdimensionen und –Facetten

Entgegengesetzt zu den handlungsorientierten Kompetenzen wie Engagement, Initiative und Leistungsmotivation liegen Kompetenzen wie Genauigkeit oder Qualitätsbewusstsein. Diese Dichotomie ist auch intuitiv gut nachvollziehbar. Alltagspsychologisch ausgedrückt: Wer schnell und zügig arbeitet, muss bei der Genauigkeit und beim Qualitätsanspruch Abstriche machen.

Auch wenn sich die Karte nicht in homogene Cluster präsentiert, so sind doch inhaltlich sinnvolle Zusammenhänge zu erkennen. Nimmt man sich bei der Analyse der Daten bereits existierende Kompetenzmodelle zu Hilfe, so lassen sich die hier abgebildeten Kompetenzbegriffe übergeordneten Kompetenzcluster zuordnen.

So erinnert die grobe inhaltliche Kompetenzstruktur zum Beispiel an die drei von Wunderer (2007) postulierten Schlüsselkompetenzen Gestaltungskompetenz, Umsetzungskompetenz und soziale Kompetenz. Zur Überprüfung dieser Hypothese wurde eine Validierungsstudie mit Prof. Wunderer (Universität St. Gallen) durchgeführt. Die 89 Kompetenzbegriffe wurden von ihm in die 3 Kategorien „Gestaltungs-, Umsetzungs- und Soziale Kompetenz“ eingeteilt. Die unten abgebildete Karte illustriert die Ergebnisse.




Stress NMDS: 0.224

Abbildung 10: Struktur der Kompetenzbegriffe nach Klassifizierung von Wunderer

Dabei ist anzumerken, dass sich gemäss Wunderer einzelne Kompetenzitems nicht eineindeutig einer Kompetenzklasse zuordnen lassen, sondern teilweise „Mischlinge“ darstellen, die inhaltlich Teile von zwei bis sogar drei seiner Kompetenzklassen beinhalten. Wunderer hat diesem Umstand durch eine Reihenfolge der Items bezüglich der zugehörigen übergeordneten Schlüsselkompetenz Rechnung getragen. So sieht Wunderer die Kompetenzfacette „Belastbarkeit“ nicht nur als Soziale Kompetenz, sondern in zweiter Linie auch als Umsetzungskompetenz, während „Selbstmanagement“ auch Komponenten der Gestaltungskompetenz beinhaltet. Bei der hier gewählten Darstellung konnte jedoch diesem Umstand nicht Rechnung getragen werden, da jedes Item eine eineindeutige Zuordnung zugewiesen bekam. Dabei wurde das jeweilige Kompetenzitem derjenigen Kategorie zugeordnet, bei welcher Wunderer den höchsten Wert (1. Rang) zugeordnet hat.

Das Problem der nicht eindeutigen Zuordnung von Kompetenzitems zu übergeordneten Kompetenzklassen haben auch Von Rosenstiel und Erpenbeck (2007) thematisiert. Sie beschreiben ihre vier Kompetenzklassen (Personale Kompetenz, Sozial-Kommunikative Kompetenz, Aktivitäts- und Handlungskompetenz, sowie Fachlich-Methodische Kompetenz) als „ontologische Klassifizierungen.“ Dabei müsse bewusst bleiben, dass die Realität nicht mechanisch, kybernetisch oder selbstorganisativ sei. Wir würden vielmehr die Mechanik, Kybernetik und Selbstorganisationstheorie dazu benutzen, um zutreffende, praktikable Modelle der Realität zu entwerfen. Im Duktus dieser Argumentation erstaunt es kaum, dass man bei faktorenanalytischer Betrachtung der Daten feststellt, dass sich die Daten der Kompetenzitems nicht durch ein paar wenige voneinander unabhängige Faktoren erklären lassen. Die Varimax rotierte Hauptkomponentenanalyse ergibt eine 21 faktorielle Lösung, die 95 % der Varianz erklären.

Die eingangs aufgestellte zweite Hypothese muss demnach abgelehnt werden. Auf Ebene der Kompetenzen ergeben sich keine klaren Kompetenzcluster. Es sind zwar inhaltliche Schwerpunkte zwischen den Kompetenzitems erkennbar, die Übergänge zwischen den übergeordneten Kompetenzklassen sind jedoch fliessend und die Zuordnung einzelner Kompetenzitems zu Kompetenzklassen somit nicht eindeutig.

Als dritten Teil der Ergebnisse wollen wir nun die zwei Perspektiven der Datenanalyse kombinieren, in dem wir die fachübergreifenden Kompetenzen in den Personenraum legen.

*Ergebnisse aus der Kombination von Personen- und Kompetenzebene*

Wir haben zeigen können, dass der Personenraum unabhängig von einzelnen Kompetenzfacetten ist und einen stabilen Raum von Kompetenzprofilen einer Managementstichprobe darstellt. Möchte man nun diesen Personenraum inhaltlich beschreiben und verstehen, sollen einzelne Kompetenzfacetten zu übergeordneten Kompetenzdimensionen zusammengefasst werden und in die Personenkarte als Regressionsgeraden gelegt werden. Dabei lassen sich unterschiedliche theoretische oder empirisch hergeleitete Kompetenzmodelle als Dimensionen in die Personenkarte legen. Zu diesem Zweck eignet sich das Verfahren des Property Fittings, welches die aggregierten Kompetenzwerte einer Dimension als Regressionsgerade in die Karte einfügt.

Abbildung 11 zeigt, wie beispielsweise die Kompetenzdimensionen von Wunderer als Skalen für das anschliessende Property Fitting berechnet wurden:

**Vorgehen Regressionsanalyse "Kompetenzdimenisionen": Modell Wunderer**

**Schritt 1**

| | Unt. D&H | Problfäh. | P&O-Fäh | Off_f-Neu | Kundorient. | etc. |
|---|---|---|---|---|---|---|
| VP 1 | 8.5 | 13 | 14.5 | 2.5 | 1.5 | |
| VP 2 | 13.5 | 7.5 | 9.5 | 8 | 13.5 | |
| VP 3 | 1.5 | 1.5 | 13 | 5 | 3 | |
| VP 4 | 10.5 | 6.5 | 2.5 | 7.5 | 11 | |
| VP 5 | 2.5 | 13.5 | 9 | 1 | 13.5 | |
| VP 6 | 12 | 10 | 15 | 1.5 | 6.5 | |
| VP 7 | 12.5 | 1.5 | 6 | 12 | 11 | |
| VP 8 | 4 | 10 | 14 | 2.5 | 6 | |
| VP 9 | 9 | 7 | 13.5 | 6.5 | 8 | |
| VP 10 | 12 | 5.5 | 8.5 | 10 | 4.5 | |
| etc. | | | | | | |

**Schritt 2**

| | **VP 1** | Gewichte | GK | UK | SK | **GL** | **UK** | **SZ** |
|---|---|---|---|---|---|---|---|---|
| Unt. D&H_K | 8.5 | 1,3,2 | 1.5 | 0.5 | 1 | 12.75 | 4.25 | 8.5 |
| Problfäh._K | 13 | 1 | 3 | 0 | 0 | 39 | 0 | 0 |
| P&O-Fäh_K | 14.5 | 1 | 3 | 0 | 0 | 43.5 | 0 | 0 |
| Off_f-Neu_K | 2.5 | 1 | 3 | 0 | 0 | 7.5 | 0 | 0 |
| Kund_or_K | 1.5 | 3,2,1 | 1.5 | 1 | 0.5 | 2.25 | 1.5 | 0.75 |
| Kommfäh_K | 4 | 3,1 | 1 | 0 | 2 | 4 | 0 | 8 |
| Teamfäh_K | 2.5 | 3 | 0 | 0 | 3 | 0 | 0 | 7.5 |
| Konflikt_K | 8.5 | 3 | 0 | 0 | 3 | 0 | 0 | 25.5 |
| Leistmot_K | 5 | 3,2 | 0 | 1 | 2 | 0 | 5 | 10 |
| | | | | | | | | |
| **Summe** | | | 13 | 2.5 | 11.5 | 8.38 | 4.30 | 5.24 |

**Schritt 3**

| | GK | UK | SK |
|---|---|---|---|
| VP01 | 8.38 | 4.30 | 5.24 |
| VP02 | 9.31 | 7.29 | 7.39 |
| VP03 | 7.00 | 8.02 | 8.80 |
| VP04 | 8.34 | 6.87 | 8.45 |
| VP05 | 8.43 | 7.28 | 8.11 |
| VP06 | 9.06 | 8.38 | 6.90 |
| VP07 | 7.89 | 8.27 | 7.91 |
| VP08 | 7.65 | 7.69 | 8.48 |
| VP09 | 8.70 | 7.05 | 8.04 |
| VP10 | 10.14 | 7.48 | 6.60 |

89 Kompetenzitems

VP = Versuchspersonen (Führungskräfte)

Rohwerte gemäss Forced-Choice-Profile

Gewichte = Einordnung der Kompetenz- items gemäss Prof. Wunderer

1 = Gestaltungskompetenz (GK)

2 = Umsetzungskompetenz (UK)

3 = Soziale Kompetenz (SK)d

Multiplikation des Rohwert und des Gewichtungsfaktor

Summe Produkte (Rowert x Gewichtung-sfaktor) geteilt durch Summe Gewichtungsfaktoren

Abbildung 11: Rechenbeispiel zur Veranschaulichung des Property-Fitting Ansatzes

In einem ersten Schritt wurde das a priori angenommene Kompetenzmodell mit den vier Kompetenzclustern Leadership, Leistungsverhalten, Soziale Kompetenz und Businessfokus und den entsprechenden Kompetenzfacetten gemäss dem oben beschriebenen Verfahren herangezogen.

Die unten abgebildete Darstellung zeigt das Property Fitting gemäss den 4 Kompetenzclustern aus dem hermeneutisch entwickelten Modell.

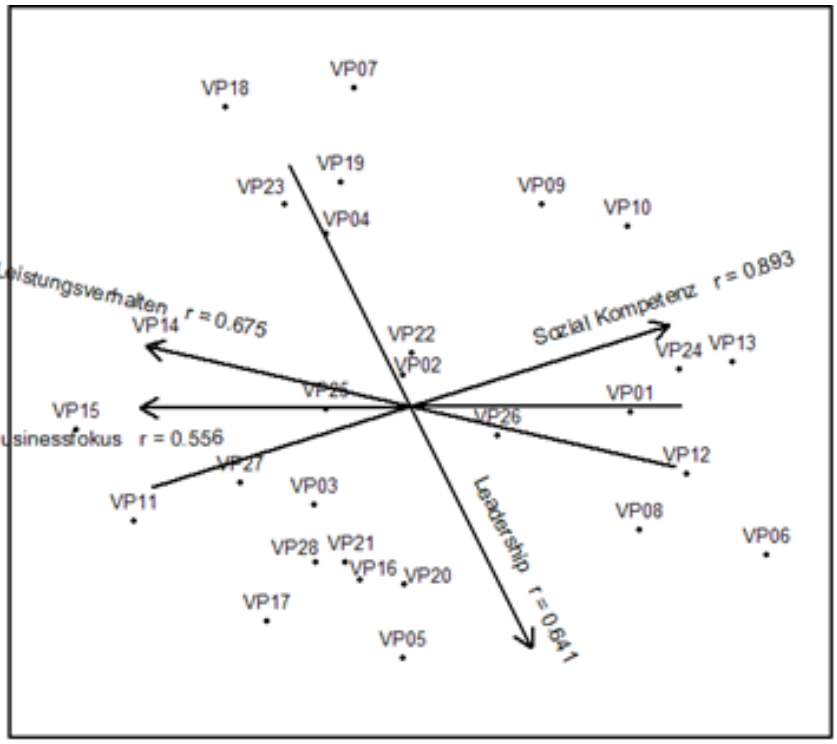



Abbildung 12: Property Fitting hermeneutisches Modell

Die Regressionskoeffizienten geben an, wie hoch der Zusammenhang zwischen der Position der Personen in der Karte und den Ausprägungsgraden auf der skalierten Kompetenzdimension ist. Bei der Höhe von Regressionskoeffizienten um 0.8 kann man davon ausgehen, dass z.B. Personen auf der rechten Seite der Karte tatsächlich den höheren Ausprägungsgrad in Sozialen Kompetenzen aufweisen als Personen auf der linken Seite der Karte. Auf der x-Achse der Karte wird eine deutliche Differenzierung in Bezug auf weiche Faktoren (Soziale Kompetenz) und härtere Faktoren (Leistungsverhalten und Businessfokus) erkennbar. Während auf der rechten Seite der Karte die sozialen kompetenten Manager abgebildet werden, sind auf der linken Seite eher die Manager zu finden, die ihre Stärken im Leistungsverhalten und Businessfokus aufweisen und umgekehrt. Auf der y-Achse kommt mittels Property Fitting das Kompetenzcluster „Leadership“ mit den Kompetenzen „Andere Inspirieren“, „Führungsverhalten“ und „Mitarbeiterförderung“ zu liegen. Im unteren Teil der Karte kommen somit Personen zu liegen, die besonders hohe Führungsqualitäten besitzen. Unerklärt bleibt in diesem Modell der obere Teil der Karte.

Auf der Suche nach Ordnung und Struktur wurde die empirisch ermittelte Itemkarte (vgl. S. 111) als Grundlage zur Kompetenzclusterung verwendet. Dabei wurden gemäss der Position der einzelnen Kompetenzbegriffe und deren inhaltlichen Bedeutung folgende 6 Konstrukte bzw. Dimensionen gebildet:

- Kommunikation
- Teamorientierung
- Genauigkeit
- Engagement
- Individuumsorientierung
- Reflexion

Die unten abgebildete Darstellung zeigt die Personenkarte und die 6 Dimensionscluster, welche wiederum mittels Property Fitting in die Karte gelegt wurden.

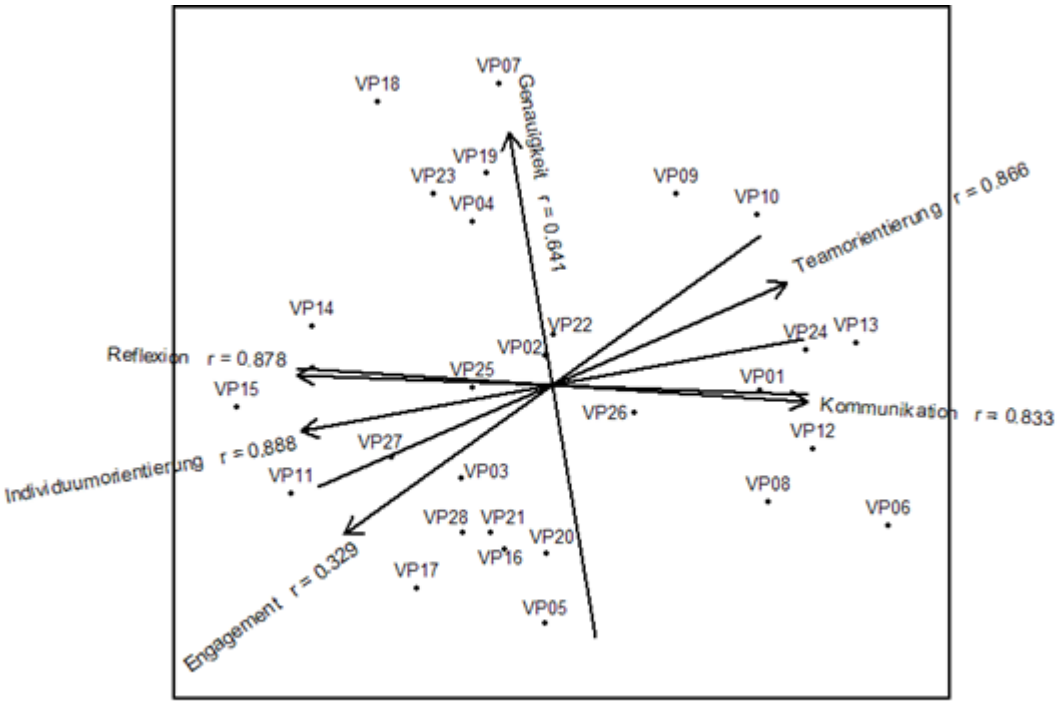



Abbildung 13: Property Fitting empirisches Modell

Richtet man den Fokus zuerst auf die rechte Seite der Karte, so fällt auf, dass gemäss diesem Modell eine Differenzierung der sozialen Kompetenz nötig ist. So lässt sich in der Karte zwei Dimensionen der sozialen Kompetenz erkennen. Einerseits die Teamorientierung bzw. Zusammenarbeit, andererseits die Kommunikationsfähigkeit.

Auf der linken Seite der Karte konnten die Cluster Reflexion, welche kognitive Aspekte beinhaltet wie Urteilsvermögen, vernetztes Denken und analytische Fähigkeiten sowie Individuumorientierung mit Kompetenzfacetten wie Selbst-bewusstsein, Durchsetzungsfähigkeit und Leistung einfordern, identifiziert werden. Die Dimension Engagement weist einen zu niedrigen Regressions-koeffizienten auf, als dass sie inhaltlich sinnvoll interpretiert werden könnte.

Von substantiellem Erklärungswert in dieser NMDS Karte ist die Dimension „Genauigkeit“, welche den oberen Teil der Karte mit einem Regressionskoeffizienten von .64 relativ gut erklärt und die Probanden auf der Y-Achse unterteilt. Im oberen Teil der Karte kommen dem zu Folge Personen zu liegen, die in Kompetenzfacetten wie Qualitätsbewusstsein, Umgang mit Ressourcen und Genauigkeit eine hohe Ausprägung aufweisen.

Neben den hermeneutischen und empirisch hergeleiteten Dimensionen (Properties), können auch Dimensionen aus gängigen Kompetenzmodellen mittels Property Fitting in die NMDS Karte gelegt werden. In dieser Studie wurde dies anhand zwei prominenter und relativ unterschiedlicher Beispiele gemacht. Zum einen wurde das bereits auf Seite 14 eingeführte Konzept des Mitunternehmertums von Wunderer (2007) verwendet, welches auf den drei Schlüsselkompetenzen Gestaltungskompetenz, Soziale Kompetenz und Umsetzungskompetenz basiert und zum anderen wurde auf das von Bartram (2001) entwickelte Kompetenzmodell „The great 8 competencies“ zurückgegriffen.

Die unten stehende Abbildung zeigt die Personenkarte und die mittels Property Fitting skalierten Kompetenzachsen nach Wunderer.

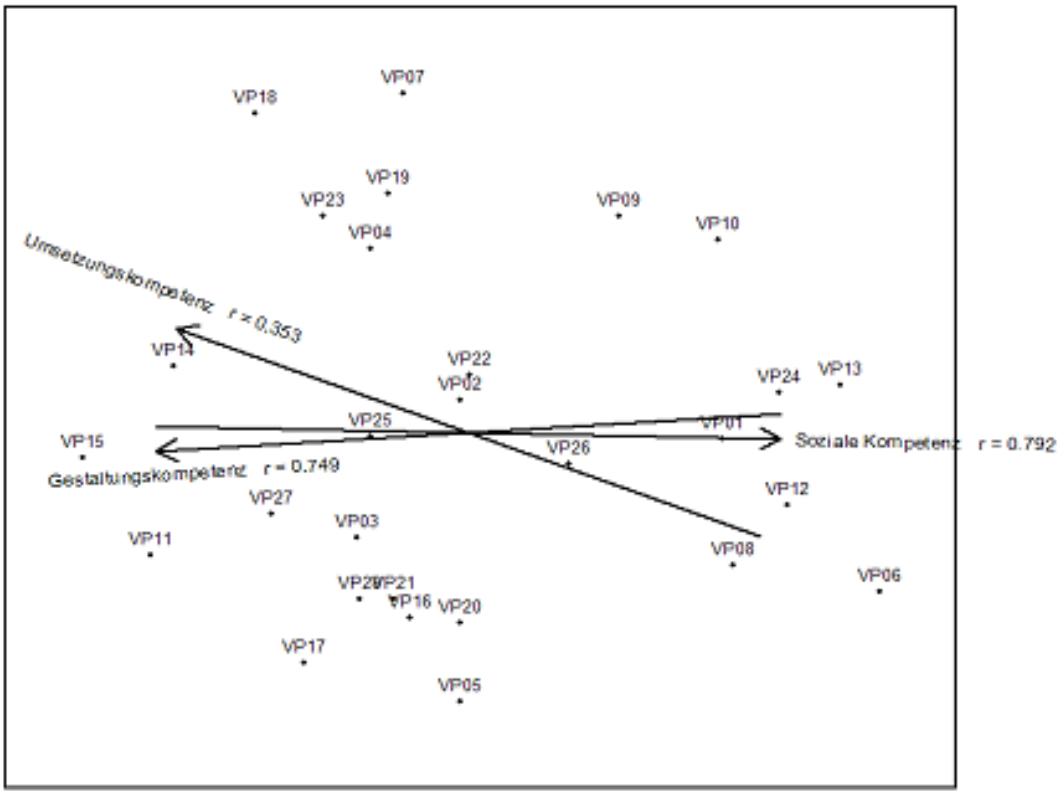



Abbildung 14: Property Fitting Wunderer Modell

Die Achsen Gestaltungskompetenz und Soziale Kompetenz lassen sich relativ gut in die Karte einpassen und widerspiegeln die Erkenntnisse aus dem Property Fitting der hermeneutischen und empirischen Karte. Die Umsetzungskompetenz lässt sich weniger gut in die Karte einpassen. Dies könnte damit zusammenhängen, dass Wunderer unter Umsetzungskompetenz etwas anderes versteht, als die hier erhobenen Daten es nahe legen.

Bei der Erklärung des Personenraums und der Beantwortung der Frage inwiefern die verschiedenen Kompetenzmodelle austauschbar sind beziehungsweise anhand welchen Modells sich die Personenkarte am besten beschreiben lässt, wurde abschliessend das Kompetenzmodell von Bartram (2001) mittels Property Fitting in die Personenkarte gelegt. Dabei wurden die 89 Kompetenzitems den 8 übergeordneten Kompetenzclustern, the great 8 competencies nach Bartram, zugeordnet. Diese Zuordnung erfolgte in Analogie an das Competency Framework von Bartram welches die 8 Kompetenzcluster in 20 Kompetenzdimensionen („Competency Dimensions“) und 112 Kompetenzkomponenten („Competency Components“) einteilt (Siehe Anhang I). Aus dieser Zuordnungsaufgabe von Kompetenzitems zu den übergeordneten Kompetenzclustern konnte wiederum mittels Property Fitting die 8

Dimensionen in die Karte gelegt werden. Die unten abgebildete Karte zeigt das entsprechende Ergebnis:

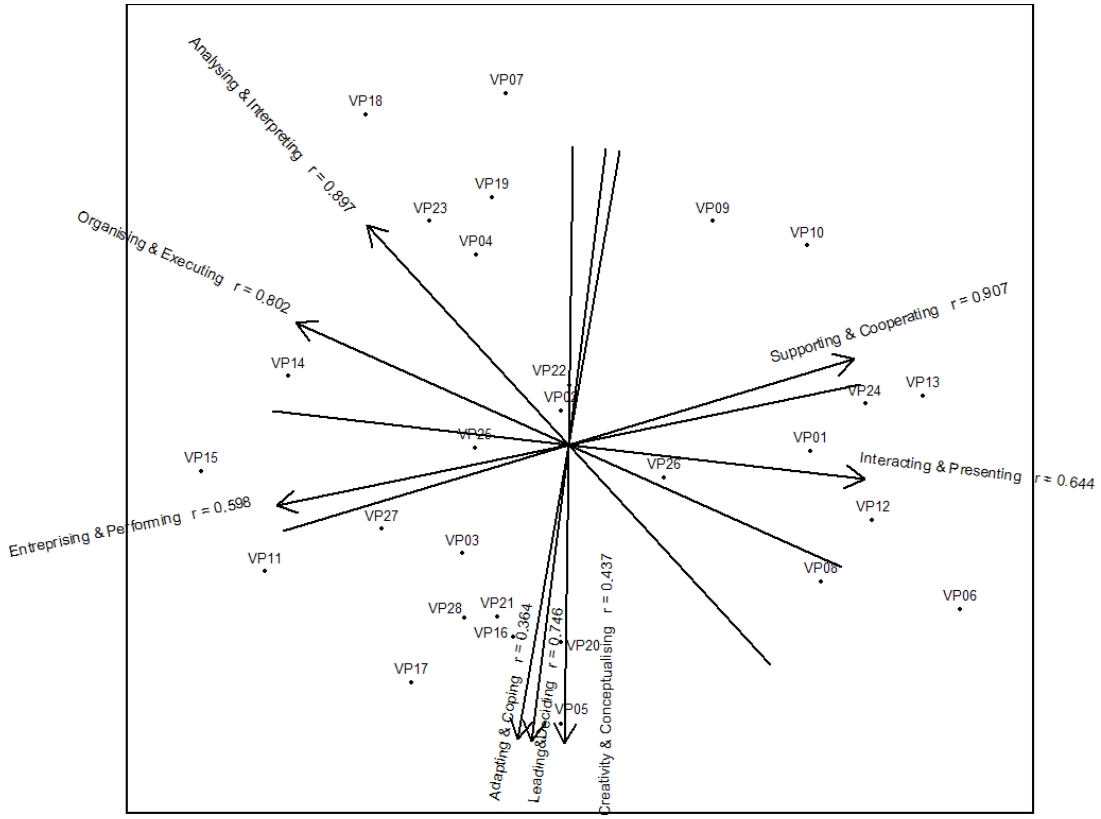



Abbildung 15: Property Fitting SHL Modell: “Great 8 Competencies”

Die Position und Richtung der Dimensionen entsprechen auch hier den in den vorherigen Karten abgebildeten Resultaten. Der zweidimensionale Raum lässt sich inhaltlich hinreichend beschreiben. Die Dimensionen lassen sich relativ gut in die Karte einpassen. Einzig die Dimensionen Adapting & Coping sowie Creativity & Conceptualising lassen sich weniger gut in die Karte legen und erklären wenig zusätzliche Varianz.

Zusammenfassend konnte gezeigt werden, dass sich der stabile Personenraum durch eine bestimmte Anzahl an Kompetenzdimensionen sinnvoll interpretieren lässt. Dabei ist vor allem der Befund interessant, dass es nicht auf das zu Grunde liegende Kompetenzmodell darauf ankommt, um die Karte der Personen sinnvoll zu beschreiben. Unabhängig von der gewählten Terminologie und Anzahl der Kompetenzfacetten, spannen die Kompetenzprofile der in dieser Studie befragten Manager einen stabilen Raum auf, der sich inhaltlich durch die Einbettung von

Regressionsgeraden erklären und interpretieren lässt. Daraus ergeben sich interessante Anwendungsmöglichkeiten für die Personalbeurteilung.

## 7.4 Diskussion

Das Ziel der vorliegenden Studie war es einerseits zu zeigen, dass mittels eines Forced-Choice-Verfahrens eine echte Alternative zu den gängigen Fragebogen Formaten der berufsbezogenen Persönlichkeitstests besteht und andererseits darzulegen, dass unabhängig von der Auswahl einzelner fachübergreifenden Kompetenzen, ein stabiler Personenraum gemessen werden kann, welcher durch übergeordnete Kompetenzdimensionen inhaltlich sinnvoll interpretiert werden kann und somit ein nützliches Personalbeurteilungsinstrument darstellt.

Aus der Literaturrecherche konnte die Erkenntnis gewonnen werden, dass die Kompetenzforschung die Tradition der Persönlichkeits- und Fähigkeitstests zur Messung des künftigen Berufserfolgs zwar nicht ablöst, aber durchaus ernstzunehmende Konkurrenz schafft. Seit dem Artikel von McClelland im Jahre 1973 erlebt die Kompetenzforschung einen regelrechten Boom. Massgebend haben die Arbeiten von Boyatzis (1982: The competent manager), über Spencer und Spencer (1993: Competence at work) sowie die auch in dieser Studie verwendeten Ergebnisse von Kurz und Bartram (2001: Competency and individual Performance: Modelling the world of work) die Kompetenzforschung geprägt. Dabei gilt es bis heute zu beachten, dass die psychologischen Konstrukte der Persönlichkeits- und Intelligenzforschung wesentlich besser validiert sind als die wenig klar abgrenzbaren Kompetenzkonstrukte, die laut Von Rosenstiel (2003, S.54) ein „Potpourri von Wissensbestandteilen, Fertigkeiten, Fähigkeiten, Qualifikationen und Persönlichkeitseigenschaften“ beinhalten. So konstatiert auch Nikolaou (2003) gleich zu Beginn seines Forschungsberichts mit folgendem Zitat, dass die Forschung in diesem doch eher neueren Feld noch grossen Nachholbedarf hat: „The assessement of work competencies and especially the development of valid and reliable measures for their assessment have attracted limited attention among the researchers in the field of personnel psychology“.

Die hier vorgestellten Ergebnisse replizieren die Erkenntnis von psychologisch wenig trennscharfen Konstrukten. So konnte anhand des Verfahrens der Nonmetrischen Multidimensionalen Skalierung gezeigt werden, dass sich bei der Darstellung der Kompetenzitems aus einem umfassenden Katalog berufsrelevanter Kompetenzen

keine eindeutige Clusterung der Kompetenzitems ergibt, sondern dass die Übergänge zwischen den einzelnen Konstrukten fliessend sind. Hier manifestiert sich ein Vorteil der Nonmetrischen Multidimensionalen Skalierung gegenüber der Faktoranalyse, welche versucht, voneinander unabhängige Konstrukte sequentiell zu extrahieren, wodurch systematische Informationen in den Daten verloren gehen.

Durch die Möglichkeit, die gesamte Varianz der Daten in einem zweidimensionalen Raum abzubilden, konnte dem Inhalt der Datenstruktur systematisch auf den Grund gegangen werden. Auch wenn keine klare Clusterung zu erkennen ist, sind die relationalen Beziehungen der Kompetenzbegriffe und deren inhaltlichen Zugehörigkeit zumindest im Sinne der Augenscheinvalidität als sinnvoll zu betrachten. So gruppieren sich inhaltlich als zusammengehörende Begriffe wie z.B. Kooperationsfähigkeit und Teamfähigkeit an einem Ort in der Karte und spannen mit anderen inhaltlich homogenen Begriffspaaren wie z.B. vernetztes Denken und Problemlösefähigkeit einen zweidimensionalen Raum auf, welcher auf eine kognitive und soziale Achse vermuten lässt. Die systematischen Zusammenhänge sind auf Itemebene jedoch nicht durchgehend zu erkennen.

Durch den Perspektivenwechsel von Item- auf Personenebene konnte schliesslich gezeigt werden, dass sich interindividuelle Unterschiede in Kompetenzprofilen durch ein Forced- Choice-Verfahren valide messen lassen. Valididät wurde dabei nicht wie in gängiger Testkonstruktion mittels faktorenanalytischer Herangehensweise der Items getestet, sondern anhand NMDS auf Personenebene. Es konnte gezeigt werden, dass die mittels Forced-Choice-Verfahren erhobenen Daten sich in einer zweidimensionaler Struktur sinnvoll abbilden lassen, sofern man die Korrelation zwischen Personen über eine Stichprobe von Merkmalen / Variablen, in diesem Fall Kompetenzitems, berechnet und relational zueinander in Beziehung setzt. Durch den Vergleich der Personen basierend auf der relationalen Ähnlichkeit von Kompetenzprofilen konnte ein stabiler Personenraum von Managertypen identifiziert werden. Die geringen Average Losses von tiefer als 0.3 der prokrusteten Karten der Split-Half-Daten untermauern die Stabilität des gemessenen Personenraums. Durch die Exploration dieses Personenraumes mittels Property Fitting konnte ein grosser Teil der Varianz durch eine bestimmte Anzahl, Position und Richtung von Regressionsgeraden erklärt werden und somit eine sinnvolle Struktur zur Beschreibung der Unterschiede in den Kompetenzprofilen der befragten

Führungskräfte gefunden werden. Basierend auf den Analysen des Property Fittings, liegt der Schluss nahe, dass unabhängig welches Kompetenzmodell man zu Grunde legt, sich ein Personenraum anhand einer bestimmten Anzahl von Kompetenzachsen beschreiben lässt. Was sich ändert ist lediglich die gewählte Semantik, z.B. „Zusammenarbeit“ oder „Supporting and Cooperating“. Der Winkel und die Richtung der Kompetenzachse bleiben jedoch erhalten. Um die Gültigkeit dieser Hypothese der Immunität der Kompetenzsemantik, d.h. die Unabhängigkeit der in der Praxis unterschiedlichen Kompetenzmodelle, könnte folgendes Versuchsdesign Aufschluss geben.

Würde man die Kompetenzprofile von Führungskräften mittels zwei oder mehrerer verschiedener Kompetenzmodelle messen, so könnte man mittels NMDS die Stabilität der Personenstruktur und somit die Unabhängigkeit der Kompetenzmodelle überprüfen. Zu diesem Zweck müssten basierend auf den Daten für beide Stichproben jeweils eine NMDS auf Personenebene berechnet werden. Mittels einer Prokrustes-Transformation, bei welcher die beiden Karten übereinander gelegt und somit die Ähnlichkeit der Karten mittels Average loss gemessen wird, könnte gezeigt werden, dass es keine grosse Rolle spielt, welches Kompetenzmodell man zur Messung von Führungskräften beizieht, solange man die verschiedenen Bereiche fachübergreifender Kompetenzen abdeckt. Das Modell von Batram (2001) scheint die verschiedenen Kompetenzbereiche von Managern gut abzudecken.

In Bezug auf mögliche Anwendungsfelder ergeben sich interessante Möglichkeiten zur effektiven Visualisierung von kompetenzbasierten Beurteilungsdaten. So können mittels der NMDS Karten die Kompetenzprofile verschiedener Personengruppen auf einen Blick erfasst und mittels Property Fitting in der Terminologie des jeweiligen Unternehmens entsprechend interpretiert werden. Durch die auf Korrelationsmassen basierende Visualisierung der Kompetenzprofile in einem zweidimensionalen Raum, kann sowohl auf Individuums- als auch auf Gruppenebene, Massnahmen in Bezug auf Stärken bzw. Entwicklungspotenziale von Mitarbeitern abgeleitet werden. Das Forced-Choice-Verfahren eignet sich dabei sowohl zur Erhebung von Selbst- als auch Fremdbeurteilungsdaten. Das Forced-Choice-Format hat den zentralen Vorteil, dass sich Urteilsverzerrungen (z.B. Akquieszenz, soziale Erwünschtheit und Impression Management) verringern bzw. teilweise sogar ganz eliminieren lassen.

Ist man jedoch an interindividuellen Niveauunterschieden interessiert, so kann im Anschluss an die Profilerhebung mittels Forced-Choice-Format noch der allgemeine

Niveauunterschied zwischen den Personen direkt abgefragt werden. (Vgl. Forschungsbericht III, Kapitl 5).

Die Entscheidung, ob ein Forced-Choice-Verfahren oder ein Fragebogeninventar für eine bestimmte empirische Fragestellung besser geeignet ist, muss einzelfallbezogen nach Abwägung der Vor- und Nachteile dieser beiden Erhebungsverfahren gefällt werden. So dient der Fragebogen der normativen Einschätzung, wie ein Merkmal, z.B. eine Kompetenzausprägung, bei einer Zielperson im Vergleich zur gesamten erhobenen Stichprobe ausgeprägt ist. Die ipsative Herangehensweise der Forced-Choice-Methode ermöglicht jedoch die differenzierte Einschätzung eines Merkmals in Bezug auf alle anderen Merkmale derselben Zielperson. Insbesondere in einem Anwendungskontext, in dem sich Personen gemäss eines bestimmten Anforderungsprofils recht einseitig darstellen wollen, ist die Verwendung eines Forced-Choice-Ansatzes besonders zu empfehlen, da durch das Forced-Choice-Format nur die Struktur, nicht jedoch der Mittelwert aller Items verändert werden kann.

Welches Verfahren gewählt werden soll, muss letztlich je nach Anwendungskontext und empirischer Zielrichtung abgewägt werden. Dem offensichtlichen Vorteil des Forced-Choice-Verfahrens, verzerrende Antworttendenzen zu unterdrücken und differenziertere Ratings zu fördern, steht der Nachteil der mangelnden Information über die Niveauunterschiede zwischen Personen gegenüber. Da diese Information jedoch nachträglich an eine Erhebung relativ einfach und effizient abzufragen ist, ist weitere Forschung im Hinblick auf die Validität und praktische Anwendbarkeit des Forced-Choice-Verfahrens wünschenswert.

## 7.5 Literatur

Batram, D. (2005): The great eight competencies : A criterion-centric approach to validation. *Journal of Applied Psychology*, 90, 6, 1185-1203.

Boyatzis, R.E. (1982). The competent manager. In A. Stewart (ed.), *Motivation and Society*. San Francisco CA.: Jossey Bass.

Boyatzis, R.E. (2006). Leadership competencies. In Ronald Burke and Cary Cooper (eds.), *Inspiring Leaders*, London: Routledge Press (Taylor & Francis Group). pp. 119-148.

Boyatzis, R., Goleman, D., & Rhee, K. (2000). Clustering competence in emotional intelligence: Insights from the emotional competence inventory (ECI). In R. Bar-On & J.D.A. Parker (eds.): *Handbook of emotional intelligence* (pp. 343-362). San Francisco: Jossey-Bass.

Dulewicz, V. & Young, M. (2008). Similarities and Differences between Leadership and Management: High-Performance Competencies in the British Royal Navy. *British Journal of Management*, 19, 1, 17-32.

Erpenbeck, J./von Rosenstiel, L. (2007). Einführung. In: Erpenbeck, J./ von Rosenstiel, L. (Hrsg.): *Handbuch Kompetenzmessung. Erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, :pädagogischen und psychologischen Praxis.* Stuttgart: Schäfer-Poeschel Verlag.

Goleman, D. (1996). Emotional Intelligence: why it can matter more than IQ. London: Bloomsbury.

Hossiep, R. & Paschen, M. (2003, unter Mitarbeit von O. Mühlhaus). Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP) (2. Aufl.) Göttingen: Hogrefe.

Kurz, R. & Bartram, D. (2001). Competency and individual performance: Modelling the world of work. Internal SHL Memorandum. Thames Ditton: SHL.

Läge, D. (2001). Ähnlichkeitsbasierte Diagnostik von Sachwissen. Habilitationsschrift, Universität Zürich, Zürich.

Matthys, A,. Zuber, T. & Läge, D. (2004): Der Zusammenhang von Anforderungs- und Kompetenzprofilen in der Management Diagnostik. Unveröffentlichte Lizentiatsarbeit, Psychologisches Institut Universität Zürich.

McClelland, D.C. (1973). Testing for competence rahter than for intelligence. *American Psychologist*, 28, 1-14.

Nikolaou, I. (2003). The Development and Validation of a measure of Generic Work Competencies. *International Journal of Testing*, 3, 309-319.

Sarges, W. (2001). Competencies statt Anforderungen – nur alter Wein in neuen Schläuchen? In Riekhof, H-C (Hrsg.), *Strategien der Personalentwicklung* (5., überarb. U. erw. Aufl.). Wiesbaden: Gabler.

Schippmann, J.S. (1999). The practice of competency modelling. *Personnel Psychology*, 53, 703-740.

Spencer, L.M., Jr. & Spencer, S.M. (1993). Competence At Work - Models For Superior Performance. New York: Wiley.

Tedeschi, J. T., & Norman, N. (1985). Social power, self-presentation, and the self. In B. R. Schlenker, (Ed.), *The self and social life* (pp. 293–322). New York: McGraw-Hill.

Spencer, L.M., Jr. & Spencer, S.M. (2008). Competence At Work - Models For Superior Performance. New York: Wiley.

Wunderer, R. (2007). Führung und Zusammenarbeit - eine unternehmerische Führungslehre. 7. überarbeitete Auflage. Köln : Luchterhand.

# Anhang I:

## Great Eight, 20 Competency Dimension and 112 Competency Component titles from the SHL Universal Competency Framework (Bartram 2005)

**1 Leading and Deciding**
1.1 Deciding & Initiating Action
1.1.1 Making Decisions
1.1.2 Taking Responsibility
1.1.3 Acting with Confidence
1.1.4 Acting on Own Initiative
1.1.5 Taking Action
1.1.6 Taking Calculated Risks
1.2 Leading and Supervising
1.2.1 Providing Direction and Coordinating Action
1.2.2 Supervising and Monitoring Behavior
1.2.3 Coaching
1.2.4 Delegating
1.2.5 Empowering Staff
1.2.6 Motivating Others
1.2.7 Developing Staff
1.2.8 Identifying and Recruiting Talent

**2 Supporting and Cooperating**
2.1 Working with People
2.1.1 Understanding Others
2.1.2 Adapting to the Team
2.1.3 Building Team Spirit
2.1.4 Recognizing and Rewarding Contributions
2.1.5 Listening
2.1.6 Consulting Others
2.1.7 Communicating Proactively
2.1.8 Showing Tolerance and Consideration
2.1.9 Showing Empathy
2.1.10 Supporting Others
2.1.11 Caring for Others
2.1.12 Developing and Communicating Self-knowledge and Insight
2.2 Adhering to Principles and Values
2.2.1 Upholding Ethics and Values
2.2.2 Acting with Integrity
2.2.3 Utilizing Diversity
2.2.4 Showing Social and Environmental Responsibility

**3 Interacting and Presenting**
3.1 Relating & Networking
3.1.1 Building Rapport
3.1.2 Networking
3.1.3 Relating Across Levels
3.1.4 Managing Conflict
3.1.5 Using Humor
3.2 Persuading and Influencing
3.2.1 Making an Impact
4.3 Analyzing
4.3.1 Analyzing and Evaluating Information
4.3.2 Testing Assumptions and Investigating
4.3.3 Producing Solutions
4.3.4 Making Judgments
4.3.5 Demonstrating Systems Thinking

**5 Creating and Conceptualizing**
5.1 Learning and Researching
5.1.1 Learning Quickly
5.1.2 Gathering Information
5.1.3 Thinking Quickly
5.1.4 Encouraging and Supporting Organizational Learning
5.1.5 Managing Knowledge
5.2 Creating and Innovating
5.2.1 Innovating
5.2.2 Seeking and Introducing Change
5.3 Formulating Strategies and Concepts
5.3.1 Thinking Broadly
5.3.2 Approaching Work Strategically
5.3.3 Setting and Developing Strategy
5.3.4 Visioning

**6 Organizing and Executing**
6.1 Planning and Organizing
6.1.1 Setting Objectives
6.1.2 Planning
6.1.3 Managing Time
6.1.4 Managing Resources
6.1.5 Monitoring Progress
6.2 Delivering Results and Meeting Customer Expectations
6.2.1 Focusing on Customer Needs and Satisfaction
6.2.2 Setting High Standards for Quality
6.2.3 Monitoring and Maintaining Quality
6.2.4 Working Systematically
6.2.5 Maintaining Quality Processes
6.2.6 Maintaining Productivity Levels
6.2.7 Driving Projects to Results
6.3 Following Instructions and Procedures
6.3.1 Following Directions
6.3.2 Following Procedures
6.3.3 Time Keeping and Attending
6.3.4 Demonstrating Commitment
6.3.5 Showing Awareness of Safety Issues
6.3.6 Complying with Legal Obligations
7 Adapting and Coping
**7.1 Adapting and Responding to Change**
7.1.1 Adapting

3.2.2 Shaping Conversations
3.2.3 Appealing to Emotions
3.2.4 Promoting Ideas
3.2.5 Negotiating
3.2.6 Gaining Agreement
3.2.7 Dealing with Political Issues
3.3 Presenting and Communicating Information
3.3.1 Speaking Fluently
3.3.2 Explaining Concepts and Opinions
3.3.3 Articulating Key Points of an Argument
3.3.4 Presenting and Public Speaking
3.3.5 Projecting Credibility
3.3.6 Responding to an Audience
**4 Analyzing and Interpreting**
4.1 Writing and Reporting
4.1.1 Writing Correctly
4.1.2 Writing Clearly and Fluently
4.1.3 Writing in an Expressive and Engaging Style
4.1.4 Targeting Communication
4.2 Applying Expertise and Technology
4.2.1 Applying Technical Expertise
4.2.2 Building Technical Expertise
4.2.3 Sharing Expertise
4.2.4 Using Technology Resources
4.2.5 Demonstrating Physical and Manual Skills
4.2.6 Demonstrating Cross Functional Awareness
4.2.7 Demonstrating Spatial Awareness
7.1.2 Accepting New Ideas
7.1.3 Adapting Interpersonal Style
7.1.4 Showing Cross-cultural Awareness
7.1.5 Dealing with Ambiguity
7.2 Coping with Pressure and Setbacks
7.2.1 Coping with Pressure
7.2.2 Showing Emotional Self-control
7.2.3 Balancing Work and Personal Life
7.2.4 Maintaining a Positive Outlook
7.2.5 Handling Criticism
**8 Enterprising and Performing**
8.1 Achieving Personal Work Goals and Objectives
8.1.1 Achieving Objectives
8.1.2 Working Energetically and Enthusiastically
8.1.3 Pursuing Self-development
8.1.4 Demonstrating Ambition
8.2 Entrepreneurial and Commercial Thinking
8.2.1 Monitoring Markets and Competitors
8.2.2 Identifying Business Opportunities
8.2.3 Demonstrating Financial Awareness
8.2.4 Controlling Costs
8.2.5 Keeping Aware of Organizational Issues

# 8 Kongruenz von Selbst- und Fremdbild bei ipsativer Messung

## 8.1 Einleitung

Wie wirklich ist die Wirklichkeit? Mit dieser oft zitierten Frage trifft Paul Watzlawik (1976) auch heute noch den Kern der multiperspektivischen Leistungsbeurteilung. Die mangelnde Kongruenz verschiedener Beurteilergruppen widerspiegelt diese Mehrdeutigkeit sozialer Wirklichkeitskonstruktion und ist ein bekannter und gut dokumentierter Effekt, welcher in zahlreichen Metaanalysen bestätigt wurde (Mabe & West, 1982; Harris & Schaubroeck, 1988; Conway & Huffcutt, 1997, Heidemeier & Moser, 2002, 2009). Die in der kürzlich veröffentlichten Metaanalyse von Heidemeier & Moser (2009) berichtete Korrelation zwischen Selbst- und Fremdurteil beträgt $r=0.22$ ($\rho=0.34$) und erhärtet noch einmal die in den bereits früher durchgeführten Metanalysen gefundenen tiefen Korrelationen.

Als ob man dem Phänomen verschiedener Wirklichkeiten beziehungsweise der Subjektivität unserer Wahrnehmung nicht recht Glauben schenken wollte, hat die Forschung zahlreiche Versuche unternommen, Wege zu finden, um die Kongruenz von verschiedenen Beurteilergruppen in der Personalbeurteilung zu erhöhen. Die Metaanalyse von Heidemeier und Moser (2009) gibt einen guten Überblick über den Effekt verschiedener Moderatorvariablen (Charakteristika der Stelle sowie der beteiligten Personen, Situationsvariablen, Skalenformat etc.), die auf den Beurteilungsprozess einwirken und die Kongruenz zwischen Selbst- und Fremdurteilen beeinflussen. Die 102 Forschungsartikel im Zeitraum von 1955 bis 2007, die in die Metaanalyse eingeflossen sind, beruhen auf zwei Gemeinsamkeiten: Erstens wurde die Beurteilerübereinstimmung zwischen Selbst- und Fremdurteilen jeweils über Korrelationen und über Mittelwertsunterschiede gemessen und zweitens wurden die den verschiedenen Studien zugrundeliegenden Daten jeweils mit einem gängigen Fragebogen erhoben, bei denen die Beurteiler eine Einschätzung zur Gesamtleistung als auch zur Ausprägung auf einzelnen Dimensionen vornehmen müssen. Sie werden also gezwungen, Gesamtprofilhöhe und Profilinformationen der Beurteilten zu vermengen.

Ziel der vorliegenden Studie ist es, diese beiden Parameter zu verändern. Zum einen sollen die Daten nicht mittels eines Fragebogens mit einer mehrstufigen Einschätzungsskala erfolgen, sondern anhand eines kompetenzbasierten Forced-Choice-Verfahrens und zum anderen soll die Übereinstimmung von Selbst- und Fremdbeurteilung nicht über die Angabe

von Korrelationskoeffizienten erfolgen, sondern mittels Nonmetrischer Multidimensionaler Skalierung (NMDS) dargestellt werden.

Bevor auf die interessierenden Fragestellungen und Methodik näher eingegangen wird, sollen theoretische Grundlagen der multiperspektivischen Beurteilungen und auf einige empirische relevante Forschung eingegangen werden.

*Methoden der Leistungsbeurteilung*

Für die Leistungsbeurteilung von Mitarbeitenden stehen verschiedene Methoden zur Verfügung. Sie kann mittels freier Eindrucksschilderung oder verschiedener Skalierungsverfahren (Auswahl-, Rangordnung- oder Einstufungsverfahren) vorgenommen werden. Dabei können auf inhaltlicher Ebene prinzipiell Eigenschaften, Fähigkeiten, Verhalten und/oder Ergebnisse betrachtet werden (für einen Überblick: Liebel & Oechsler, 1994; Schuler, 2003).

Wie bereits zu Beginn bei der zitierten Metaanalyse von Heidemeier & Moser (2009) angedeutet, werden mit Abstand am häufigsten *Fragebogen* zur Messung von Führungsverhalten eingesetzt, wobei Fragebögen zu den Einstufungsverfahren zählen. Der wohl gewichtigste Vorteil aus multiperspektivischer Betrachtungsweise ist, dass auf verhältnismässig ökonomische Art und Weise Fremd- und Selbstbeschreibungen kombiniert werden können. Dazu muss der Fragebogen lediglich von der Führungskraft selbst und zusätzlich von weiteren Personen ausgefüllt werden. Ziel kann sein, einzelne Führungskräfte in ihrem Führungsverhalten zu beurteilen (Individualanalyse) oder einen Trend quer über das Unternehmen zu gewinnen (Klimaanalyse). Analog zu den unterschiedlichen Zielebenen kann auch die Analyseebene mehr oder weniger summarisch sein: Ein allumfassender Führungsscore, einzelne Führungsdimensionen oder einzelne Iteminhalte können im Vordergrund stehen. Die verfügbaren Instrumente unterscheiden sich weniger der Erhebungsform nach (es sind zumeist Fragebögen mit Einzel-Items und graduellen Zustimmungsskalen) als vielmehr hinsichtlich der gemessenen Inhalte (also unterschiedlicher „Dimensionen“ oder „Faktoren“). Insgesamt erscheint der Markt an Fragebögen für Führungsfragebogen sehr unübersichtlich, da eine Vielzahl an Instrumenten wissenschaftlich unveröffentlicht blieb. Drei Fragebogeninstrumente scheinen uns aus der Praxis heraus besonders erwähnenswert: Der Fragebogen zur Vorgesetzten-Verhaltens-Beschreibung (Fittkau & Fittkau-Garthe, 1971), der die deutschsprachige Führungsforschung lange Zeit prägte, die Qualitative Führungsstilanalyse (Fennekels, 2000), als ein aktuelleres Instrument der Praxis, und der Multifactor Leadership Questionnaire (Bass, 1985), der ein besonders breites Spektrum an Führungsverhalten erfasst („Full Range of Leadership“).

Neben den hier beschriebenen Fragebögen, werden auch Verfahren eingesetzt, welche auf einem Forced-Choice-Format beruhen und zu den ipsativen Verfahren gehören. Bei ipsativen Verfahren bleibt die Summe der eingeschätzten Leistung bzw. Kompetenzausprägung konstant, d.h. die Profilhöhe der verschiedenen Beurteilten wird nicht erhoben, sondern lediglich das relative Profil in den einzelnen Dimensionen. Auch wenn dieses Verfahren weniger häufig zum Einsatz kommt, so haben Forced-Choice-Verfahren den Vorteil, dass sie nicht den bekannten Urteilstendenzen (Milde-/Strengetendenz, Tendenz zur Mitte) unterliegen. (Heggestad, McCloy & Reeve, 2006; Christiansen, Burn & Montgomery, 2006). Zudem können, und das ist uns wichtiger, mittels Forced-Choice-Format differenziertere Profilinformationen gewonnen werden als im Fragebogenformat. Der Grund ist, dass im Likert-Skalen basierten Fragebogenformat Informationen über Profilhöhe und Profil vermischt werden und dies den Beurteiler nach unserer Einschätzung kognitiv überfordert. Schliesslich muss er sowohl das Level der absoluten Profilhöhe (im Vergleich zu anderen Personen) einhalten als auch die unterschiedliche Ausprägung in einzelnen Dimensionen im Auge behalten (Vgl. hierzu Forschungsbericht III).

*Stand der Forschung zur interperspektivischen Passung der Beurteilungen von beruflicher Leistung*

In Analogie zur Verbreitung der gängigen mehrstufigen Fragebögen in der Praxis, hat sich auch die Forschung bei der Überprüfung der Beurteilerübereinstimmung mehrheitlich auf traditionelle Fragebogendaten und nicht auf Forced-Choice-Verfahren gestützt. Dabei wurde der Fokus sowohl auf Niveauunterschiede als auch auf korrelative Übereinstimmung gelegt.
Christansen, Burn & Montgomery (2005) machen vor allem die psychometrischen Probleme aufgrund der Abhängigkeit von Forced-Choice-Skalen für das mangelnde Forschungsinteresse an Forced-Choice-Daten verantwortlich.
Betrachtet man die Ergebnisse der traditionellen Fragebogenforschung, so stellt man fest, dass sich beim Vergleich von Selbst- und Fremdbild erhebliche Unterschiede ergeben. Man kann sagen: viele Wirklichkeiten – je nach Perspektive und individuellen Motiven fällt die Bewertung des Leistungsverhaltens einer Person anders aus.

*Niveauunterschiede*

Ein fundamentales Problem aller gängigen Fragebogenformate, welche sowohl Profilhöhe als auch das Profil messen, sind die Niveauunterschiede unterschiedlicher Beurteiler. Führungskräfte beschreiben sich selbst (unter anderem ihr Führungsverhalten) im Durchschnitt deutlich *besser* als sie von anderen beschrieben werden. Das bedeutet jedoch

nicht automatisch, dass die Selbstbeschreibung immer fehlerhaft und die Fremdbeschreibung immer richtig ist – oder umgekehrt. Es handelt sich zunächst um verschiedene Sichtweisen des gleichen Verhaltens.

Wie in der Einleitung angetönt, haben verschiedene Meta-Analysen das Ausmass der Diskrepanz zwischen Selbst- und Fremdurteilen untersucht. Es muss darauf verwiesen werden, dass die periodisch vorgelegten Meta-Analysen Studien eingehen, die die multiperspektivische Beurteilung *verschiedenster* Verhaltensbereiche erfassen. So sind die Fokuspersonen der in die Meta-Analysen eingehenden Studien nicht nur Führungskräfte aus Industrie, Dienstleistung und Militär, sondern auch technische Angestellte, Arbeiter, Pfarrer, Lehrer und andere (vgl. Tabelle 1). Mabe und West (1982) legen sich nicht einmal auf berufliche Leistung fest, sondern integrieren beispielsweise auch schulische Leistung oder im Labor provoziertes Leistungsverhalten unterschiedlicher Art. Das bedeutet, dass die Analysen zwar die für die vorliegende Studie relevante Stichprobe der Führungskräfte umfassen, sich aber bei weitem nicht auf sie beschränken.

| | Mabe & West, 1982 | Harris & Schaubroeck, 1988 | Conway & Huffcutt, 1997 | Heidemeier & Moser, 2002 | Heidemeier & Moser, 2009 |
|---|---|---|---|---|---|
| ρ Selbst – Ab | .29<br>*k* = 55 | .22 (.35)<br>*k* = 36<br>.27[14]<br>*k* = 11 | .22 (.31)<br>*k* = 50<br>.19 (.29)[b]<br>*k* = 19 | .22 (.35)<br>*k* = 76<br>.19 (.32)[b]<br>*k* = 21 | .22(.34)<br>*k*=115 |
| ρ Selbst – Auf | - | - | .14 (.26)<br>*k* = 50 | - | - |
| ρ Ab - Auf | - | - | .22 (.57)<br>*k* = 22 | - | - |
| *d* Selbst - Auf | - | .70<br>*k* = 18 | | .35<br>*k* = 54<br>.34b<br>*k* = 17 | .32<br>*k*=89 |

| | | | | | |
|---|---|---|---|---|---|
| Stichprobe | 81% Studierende (nicht nur berufliche Leistung), Rest: u.a. Management | Verschiedene Bereiche, z.B. Management, Service | Verschiedene Dimensionen der Arbeitsleistung | Feldstudien zu Arbeitsleistung in aktuellem Berufs-verhältnis, versch. Sektoren, 72% Industrie | Feldstudien im Zeitraum von 1955 bis 2007 |
| Aussenkriterium | Verschiedene, z.B. Abwärts-ratings (16%) oder objektive Tests (50%) | Abwärtsratings- und Peerratings[c] | Abwärts- und Aufwärtsratings | Abwärtsratings | Abwärtsratings |

*Anmerkungen.*
*k* = Anzahl der unabhängigen Stichproben
In Klammern sind die korrigierten Indizes angegeben. Bei allen korrigierten Indizes wird die Reliabilität der Ratings berücksichtigt. Bei Harris und Schaubroeck (1988) wurden sie zusätzlich um die Einschränkung des Ranges bei Selbstbeurteilungen (aufgrund der Mildetendenz) korrigiert.
[a] Die berichteten Korrelationen beschreiben die Zusammenhänge zwischen der Selbstbeschreibung und Aussenkriterien, die nur zu 16% Abwärtsbeurteilungen darstellen (vgl. entsprechende Spalte der Tabelle).
[b] Ergebnis für Teilstichprobe "Management/Führung". Bei allen Angaben, die Aufwärtsbeurteilungen enthalten, ist naturgemäss die Beurteilung von Führungskräften betroffen, da andere Personenstichproben keine Mitarbeiter haben.
[c] Die Ergebnisse der Peerratings sind in der Tabelle nicht berücksichtigt.

Besonders hoch erweisen sich die Niveau-Unterschiede zwischen Selbst- und Fremdbildern: So fällt die Selbstbeurteilung deutlich besser aus als die Abwärtsbeurteilung ($d$ = .70 bei Harris & Schaubroeck, 1988; $d$ = .35 bzw. $d$ = .32 bei Heidemeier & Moser, 2002 und 2009). Heidemeier und Moser (2009) geben auch für die Management-Stichprobe selbst einen Unterschied von einer Drittel Standardabweichung an. Weniger überhöht erscheint das Selbstbild im Vergleich zur Kollegenbeschreibung ($d$ = .28 bei Harris & Schaubroeck, 1988). Da in keine Meta-Analyse Aufwärtsbeschreibungen einbezogen wurden, muss für entsprechende Aussagen gänzlich auf andere Studien zurückgegriffen werden. Mount und Scullen (2001) zitieren zwei nicht publizierte Studien, die an grossen Manager-Stichproben alle vier Perspektiven hinsichtlich der Niveau-Unterschiede verglichen, und halten fest, dass sich Aufwärtsbeschreibungen generell kaum im Niveau von Kollegen- und Vorgesetztenbeschreibungen unterscheiden (auf einzelnen Dimensionen leicht über oder unter diesen liegen) und sich folglich ebenfalls deutlich von Selbstratings unterscheiden. Wird zur Führungsbeschreibung der MLQ herangezogen, dann liegt die Selbstbeschreibung um etwa einen halben Skalenpunkt (bei einer 5er Skala) höher als die Aufwärtsbeschreibung (Atwater & Yammarino, 1992; Bass & Yammarino, 1991). Insgesamt fallen Selbstbeschreibungen also

deutlich besser aus als Fremdbeschreibungen von Vorgesetzten, Kollegen und Mitarbeitern, während sich die letzten drei kaum unterscheiden (Fleenor, McCauley & Brutus, 1996). Auch im Vergleich zu externen Beratern überschätzen Manager ihre Management-Fähigkeiten (Furnham & Stringfield, 1998).

Nicht zu vergessen ist selbstverständlich, dass es auch Führungskräfte gibt, die sich in Übereinstimmung mit Fremdbeurteilern befinden oder sich negativer als diese bewerten. Tornow (1993) berichtet in seinem Diskussionspapier folgende Zahlen: Nur zehn Prozent der Manager schätzen sich im Einklang mit ihren Mitarbeitern ein, die anderen weichen mindestens um eine halbe Standardabweichung vom Fremdbild ab. Circa zwei Drittel der diskrepanten Einschätzungen sind dergestalt, dass sich die Führungskräfte besser beschreiben als ihr Fremdbild. Erstaunlich sind andere Zahlen, die annähernd auf eine Drittelung in Unterschätzer, Übereinstimmer und Überschätzer hinweisen (Atwater, Roush & Fischthal, 1995, Van Velsor, Taylor & Leslie, 1993). Möglicherweise ist ein Teil dieses offensichtlichen Widerspruchs dadurch aufzuklären, dass das Ausmass der Überschätzung das der Unterschätzung deutlich übersteigt, so dass im Durchschnitt eine Überschätzung resultiert. Wichtig sind erste Hinweise darauf, dass die Diskrepanz nicht durch eine Perspektive allein verursacht wird, sondern sowohl durch die Selbst- als auch Fremdperspektive. Bei einer hohen Diskrepanz scheint also nicht allein die Selbstperspektive nach oben zu tendieren, sondern auch die Fremdperspektive nach unten. Für die Unterschätzer zeichnet sich ein überraschendes Bild ab: Sie schätzen sich selbst im Vergleich zu anderen am schlechtesten ein, werden von anderen jedoch am besten beurteilt (Atwater & Yammarino, 1992; Van Velsor, Taylor & Leslie, 1993).

*Korrelativer Zusammenhang*

Der in diesen letzten Abschnitten wiederholt berichtete und bereits in die Lehrbücher eingegangene Befund für den diskrepanten Selbst-Fremd-Vergleich wäre weniger bedeutend, wenn er durch Addition einer Konstanten aufgelöst werden könnte. Dann würden die Profile der Selbst- und Fremdbeschreibung trotz der Niveau-Unterschiede übereinstimmen, also hoch korrelieren. Die Mittelwertsdifferenzen könnten beispielsweise mit einem Ankereffekt (Tversky & Kahneman, 1974) erklärt werden. Doch Führungskräfte beschreiben ihr eigenes Führungsverhalten qualitativ *anders* als Aussenstehende. So finden sich auch auf korrelativer Ebene nur geringe Zusammenhänge zwischen dem Selbstbild der Führungskraft und den Fremdbildern. Dies ist ein gut untersuchtes Phänomen mit erstaunlich konsistenter Befundlage. Für allgemeine berufliche Leistungsbeurteilungen schätzen die Meta-Analysen

folgende Zusammenhänge mit dem Abwärts-/Kollegenbild: ρ = .22 (Conway & Huffcutt, 1997), ρ = .22 (Harris & Schaubroeck, 1988) beziehungsweise ρ = .22 (Heidemeier & Moser, 2002). Noch geringere Korrelationen ergeben sich für die Substichprobe der Führungskräfte: ρ = .19 (Conway & Huffcutt, 1997) beziehungsweise ρ = .08 (Mabe & West, 1982; nicht nur Abwärtsbeurteilung, sondern verschiedene Aussenkriterien). Von allen 12 Leistungsbereichen, die Mabe und West (1982) untersuchten, fallen die Korrelationen zwischen Selbstbewertung und Kriterium im Management-Bereich damit am geringsten aus. In der grossen Management-Stichprobe (*N* = 2.056) von Fleenor, McCauley und Brutus (1996) korreliert die Selbstbeschreibung der Management-Kompetenzen mit der Abwärtsbeschreibung etwas geringer (*r* = .19) als mit der Aufwärtsbeschreibung (*r* = .24), wobei die Autoren festhalten, dass die Ergebnisse für Selbst-Abwärts-Zusammenhänge stark schwanken und in der Regel höher liegen als die für Selbst-Aufwärts-Zusammenhänge. Diese erweisen sich im Unterschied zu den Selbst-Abwärts-Zusammenhängen als recht stabil. Auch Atwater, Ostroff, Yammarino und Fleenor (1998) finden für die Führungsbeschreibung von 1.326 Managern eine Korrelation von *r* = .25 zwischen Selbst- und Aufwärtsbeurteilung. In diese Grössenordnung reihen sich andere Ergebnisse widerspruchslos ein: Selbst-Aufwärts (z.B. Atwater & Yammarino, 1992 [*r* = .19 für Marinestudenten bzw. *r* = .12 für Marineoffiziere]; Baril, Ayman & Palmiter, 1994 [*r* = .16 bis .34 für unterschiedliche Beurteilungsdimensionen] Church, 1997 [*r* = .25]; London & Wohlers, 1991 [*r* = .24]). Ebenso verhält es sich in der – über den Führungsbereich hinausreichenden – Meta-Analyse von Conway und Huffcutt (1997; ρ = .14, korrigiert ρ = .26).[15]

Insgesamt scheint zu gelten, dass die Korrelationen etwas steigen, wenn das Mittel aus mehreren Fremdbeurteilungen (auch über die Perspektiven hinweg) anstatt individueller Werte mit der Selbstbewertung korreliert wird (Wohlers & London, 1989). Ausserdem bestehen höhere korrelative Zusammenhänge, wenn nur Fremdbilder miteinander verglichen

[15] Geringe interperspektivische Korrelationen, aber auch hohe Niveau-Unterschiede (Selbst > Fremd) zwischen Selbst- und Fremdbeschreibungen wurden neben dem Führungsbereich in vielen anderen Bereichen gezeigt. Dazu gehören die Beurteilung von Leistungen in Assessment-Centern (Drees, 1994; Stempfle, Hagmayer, Hübner, Iwanoff & Kaufmann, 2004: geringe Korrelation zwischen Selbst- und Beobachter-Beurteilung), klerikale Arbeit (z.B. Parker, Taylor, Barrett & Martens, 1959: hohe Niveau-Unterschiede und geringe Korrelationen), Pflegefertigkeiten (z.B. Klimoski & London, 1974: geringe Selbst Fremd-Korrelationen) oder Depressivität und Ängstlichkeit (Okazaki, 2002: von nahestehenden Personen unterschätzt und lediglich mittlere Selbst-Fremd-Korrelationen).

werden und das Selbstbild aussen vor bleibt. Diesen Befund erbringen sowohl die Meta-Analysen für die Führungsbeurteilung von Führungskräften (Conway & Huffcutt, 1997: Aufwärts-Abwärts: $\rho = .57$, Aufwärts-Peer: $\rho = .66$, Abwärts-Peer: $\rho = .79$ bzw. Harris & Schaubroeck, 1988: Abwärts-Peers: $\rho = .62$) als auch Einzelstudien (Fleenor et al. [1996]: $r = .41$ für Aufwärts-Abwärts-Korrelation bzw. $r = .24/.19$ für Selbst-Aufwärts/Abwärts-Korrelation; Furnham & Stringfield [1998]: $r = .58$ als mittlere Korrelation aus 72 Fremd-Fremd-Korrelationen im Vergleich zu $r = .13$ aus Selbst-Fremd-Korrelationen[16]). Das kann als Zeichen dafür gewertet werden, dass die Selbstbewertung einen grossen Anteil an der mangelhaften Übereinstimmung trägt. Ausserdem gibt dies einen Hinweis darauf, dass Fremdbeobachter Ähnlichkeiten zueinander aufweisen, dass aber die Selbstwahrnehmung der Führungskräfte nahezu unabhängig von den Fremdurteilen ist.

Vergleicht man die beiden berichteten Forschungsperspektiven, so stellt man unschwer fest, dass in der Forschung deutlich mehr Wert auf korrelative Zusammenhänge zwischen den Perspektiven als auf Niveau-Unterschiede gelegt wurde. Methodische Fragestellungen, insbesondere zur Reliabilität von Selbstbeurteilungen, standen dabei im Vordergrund. Dies gilt nicht nur für berufliche Leistungsbeurteilung und Beurteilung von Führung im Allgemeinen, sondern auch für spezifische Anforderungen, die an Führungskräfte gestellt werden (z.B. eigenverantwortliches Handeln, Koch, 2001). Es stellte sich die Frage, wie stabil solch diskrepante Beurteilungen über die Zeit hinweg sind. Hier stehen die Studien, die nach Veränderungsmöglichkeiten durch Feedback suchen, denjenigen gegenüber, die eine Retest-Reliabilität maximieren wollen und Werte von bis zu $r = .81$ erbringen (Nilsen & Campbell, 1993; Smither, London, Vasilopoulos, Reilly, Millsap & Salvemini, 1995). Mit dem letztgenannten Wert entspricht der Stabilitätsindex für Führungsdifferenzen in etwa dem für Abwärtsbeschreibungen von genereller Arbeitsleistung ($\rho = .81$, Meta-Analyse Viswesvaran, Ones & Schmidt, 1996). Dies erstaunt, wenn man geäusserte Kritik an der Reliabilität von Differenzscores ernst nimmt (Johns, 1981, vgl. Kapitel 5.2.1).

Man kann also aus empirischen Befunden festhalten, dass Führungskräfte ihr eigenes Leistungsverhalten im Durchschnitt besser einschätzen als dies von anderen eingeschätzt wird (Niveau-Unterschiede), dass die Selbstbeschreibungen weniger mit Fremdbeschreibungen korrelieren als diese untereinander und dass Fremdbeschreibungen stärker mit (organisationalen) Erfolgskriterien korrelieren als Selbstbeschreibungen.

---

[16] Die Fremdbeurteilungen verteilen sich auf andere Manager, Peers und Berater, wobei Peers mit den Beratern am geringsten korrelieren, während sich andere Fremd-Fremd-Korrelationen nicht unterscheiden.

Der Forschung scheint dabei der Befund am wichtigsten, dass die Selbstbeurteilung die geringsten Korrelationen zu anderen Perspektiven aufweist. Dies ist am intensivsten untersucht worden, was sich schon daran zeigt, dass diese Frage in fast alle multiperspektivischen Beurteilungsstudien integriert ist. So konnte sie auch in alle fünf beschriebenen Meta-Analysen aufgenommen werden.

In der Praxis musste man sich folglich entscheiden und bewertete lange Zeit die Abwärtsbeurteilung als Perspektive mit höchster Relevanz. Da dies die akzeptierteste Form der Leistungsbeurteilung darstellte und darstellt, wurde auch sie in vielen Studien und letztlich ebenso in allen fünf Meta-Analysen berücksichtigt. Im Unterschied dazu setzte sich die Aufwärtsbeurteilung erst in den vergangenen Jahren verstärkt durch. Obwohl alle Perspektiven für sich genommen – mehr (z.B. Abwärtsbeurteilung) oder weniger (Selbstbeurteilung) – valide sind, gewinnt die multiperspektivische Leistungsbeschreibung deutlich durch die inkrementelle Validität der verschiedenen Perspektiven. Die Aussagekraft eines vielfältigen Gesamtbildes wird nach der Meinung mancher Autoren dadurch ermöglicht, dass die Perspektiven untereinander nur gering korrelieren und folglich wenig redundante, sondern spezifische, kriteriumsrelevante Information beisteuern (Meta-Analyse von Conway, Lombardo & Sanders, 2001; Scullen, Mount & Goff, 2000). Aber macht man, wenn man das so sieht, nur aus der Not eine Tugend?

*Fragestellungen und Hypothesen*

Wie eingehend dargestellt wurde, handelt es sich bei der geringen Selbst-Fremd-Übereinstimmung (Niveau-Unterschiede und Korrelationen) um ein vielfach bestätigtes Phänomen. Die überwiegende Mehrheit der im letzten Abschnitt berichteten Ergebnisse stammt aus Studien, bei denen die Daten aus den gängigen Fragebogenformaten gewonnen wurden. Dabei wurde in den jeweiligen Metaanalysen vornehmlich die Frage nach der Interrater-Reliabilität innerhalb einer Gruppe von Beurteilern sowie die Korrelationen zwischen verschiedenen Beurteilergruppen (Vorgesetzten-, Peer-, Aufwärts- und Selbstbeurteilungen) betrachtet.

Wie in der Einleitung bereits erörtert, werden wir in dieser Studie nur zwei wesentliche Merkmale verändern: Erstens werden wir die Daten nicht mittels den gängigen Likert-skala basierten Fragebogenformaten erheben, sondern mittels Forced-Choice-Format, und zweitens wollen wir die Übereinstimmung zwischen Fremdurteilen und Selbsturteilen bzw. zwischen Fremdurteilern untereinander nicht mittels gemittelten Korrelationskoeffizienten bestimmen, sondern durch Differenzsummen, und diese werden wir mittels Nonmetrischer Multidimensionaler Skalierung auswerten. Wir gehen davon aus, dass es zur Modellierung

von Übereinstimmung zwischen verschiedenen Beurteilern, geeignetere Masse als den Mittelwert von Korrelationskoeffizienten gibt, da hohe Korrelationen zwischen einzelnen Beurteilerpaaren durch niedrige Korrelationen anderer Beurteilerpaare bei der Mittelwertbildung wieder aufgehoben werden. Das Verfahren der NMDS hat nun genau Vorteil, dass die Übereinstimmung nicht jeweils auf ein Beurteilerpaar reduziert ist, sondern die Kovarianzen aller Beurteiler zu einer Person direkt (d.h. ohne Umweg über Hauptkomponenten) und in gleicher Weise bei der Strukturbildung berücksichtigt werden. Dies erlaubt die differenzierte Auseinandersetzung der Interraterreliabilität über verschiedene Beurteiler über mehrere Beurteilte.

Wenn man an Beurteilerdaten herangeht, interessiert in erster Linie, inwiefern sich die eingeschätzten Profile zwischen Individuen tatsächlich unterscheiden oder inwiefern sie das Produkt eines Messfehlers sind. Bei der Messung der interindividuellen Profilunterschiede ergeben sich verschiedene Varianzquellen. Neben tatsächlich vorliegenden Unterschieden zwischen Personen, können diese im Messinstrument liegen, in der verzerrten, subjektiven Wahrnehmung auf Seiten der Beurteiler als auch in der Selbstwahrnehmung der Beurteilten. Wir postulieren anhand euklidischen, zweidimensionalen Karten drei Separierungsmodelle, mit denen sich drei Möglichkeiten der Realität abbilden lassen. Darauf aufbauend werden verschiedene Hypothesen abgeleitet, welche den verschiedenen Phänomenen der Personalbeurteilung auf den Grund gehen.

Das erste Modell basiert auf der radikalen Annahme, dass basierend auf gängigen Führungskompetenzen ganz spezifische Kompetenzprofile bzw. Persönlichkeitstypen gemessen werden können, die sowohl im Selbst- als auch im Fremdbild absolut kongruent sind. Alle an der Messung beteiligten Personen haben ein stimmiges und einheitliches Bild über die jeweiligen Kompetenzunterschiede verschiedener Personen.

Würde man solche Daten messen und die Profile beziehungsweise die Differenzsummen in einer zweidimensionalen euklidischen Karte abbilden, so ergäbe sich folgendes Modell, welches in Abbildung 1 dargestellt ist. Dabei kommen ähnliche Profile als Punkte im zweidimensionalen Raum sehr nahe beieinander zu liegen, während unähnliche Profile weit voneinander entfernt abgebildet werden.

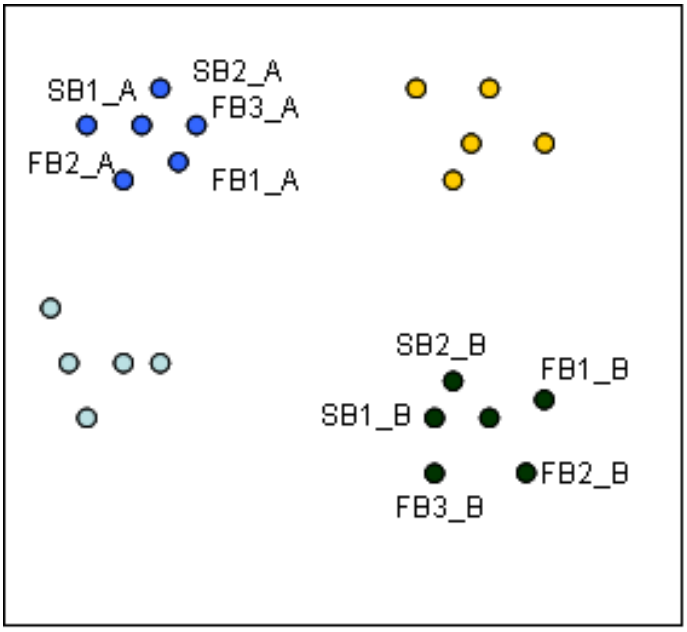



Abbildung 1: Modell: Die Punkte repräsentieren Kompetenzprofile von Personen, welche aufgrund ihrer Ähnlichkeitsbeziehungen untereinander in einem zweidimensionalen Raum abgebildet sind.
SB1 und SB2 = Selbstbilder zu zwei verschiedenen Messzeitpunkten
FB 1_A. FB2_A etc. = verschiedene Fremdbilder zu jeweils einer Person

Die entgegengesetzte, nicht minder radikale Annahme, beruht auf einer absoluten Inkongruenz von Kompetenzprofilen bzw. auf einer kompletten Unabhängigkeit des Selbstbildes und den verschiedenen Fremdbildern zu einer jeweiligen Person. Wir bezeichnen diesen Fall als Null-Modell bei dem keine Zusammenhänge zwischen Beurteilern und Beurteilten bestehen. Diese Annahme der vollkommenen Unabhängigkeit zwischen Selbst- und Fremdbildern bzw. zwischen Fremd- und Fremdbild wird in Abbildung 2 dargestellt:

.

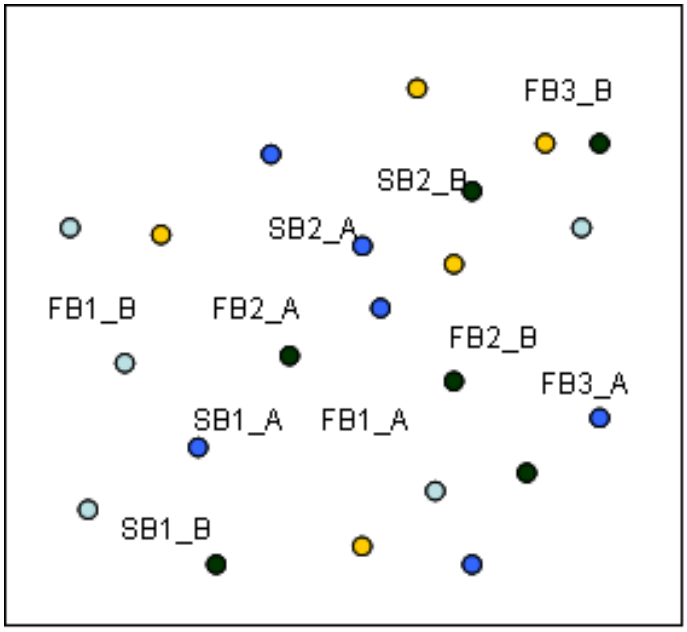



Abbildung 2: Modell: Selbst- und Fremdbilder zu einzelnen Personen stehen in keinem system. Zusammenhang

Das reale Messergebnis wird vermutlich irgendwo zwischen diesen beiden skizzierten Extrem- Modellen liegen, also nicht alle Varianz auf tatsächliche Unterschiede zwischen Personen zurückführen können, aber doch solche Unterschiede zumindest graduell finden. Dieses Modell geht davon aus, dass es durchaus ganz bestimmte Typen von Kompetenzprofilen bzw. Persönlichkeitstypen gibt, und dass diese von den meisten Beurteilern auch als solche erkannt und bewertet werden. Es werden jedoch immer auch idiosynkratische „Einzelsichten“ konzediert, die aufgrund eines anderen Erlebens bzw. unterschiedlicher Erfahrungen mit den Rollen einer Person eine komplett andere Sicht einer Person haben. Diese idiosynkratische Sicht passt dann nicht in das Bild, das die meisten Menschen von einer bestimmten Person haben. Bildet man diese Annahme in einer zweidimensionalen euklidischen Karte ab, so ergibt sich ein Modell wie in Abbildung 3 skizziert.

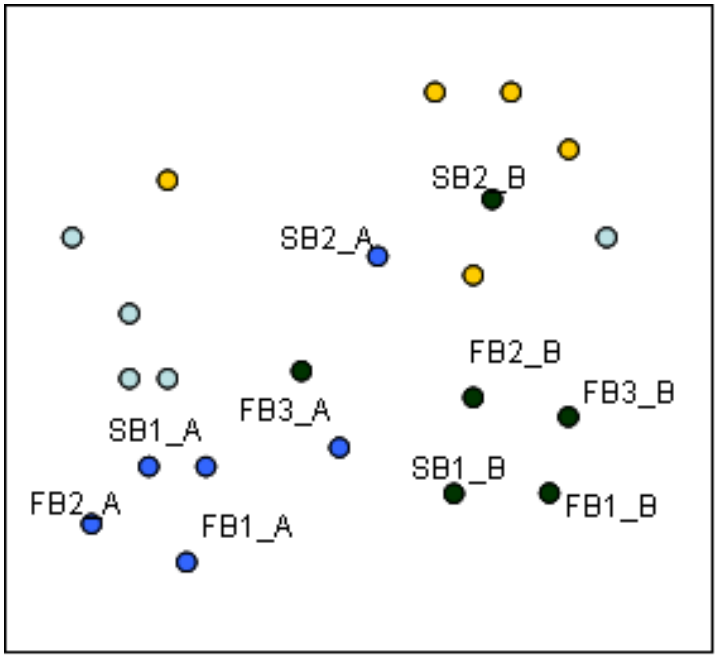



Abbildung 3: Modell: Sozialwissenschaftliche Realität bei Leistungsbeurteilungsdaten.

Die Hypothese lautet somit, dass verschiedene Beurteiler, unabhängig von Hierarchiestufe, Funktion und Unternehmenskontext, mehrheitlich eine kongruente Meinung über eine bestimmte Person haben. Formen wir das schwammige Wort „mehrheitlich“ in eine inferenzstatistische Aussage um, so ergeben sich folgende operationalisierte Hypothesen:

H1a: Die Intraclusterdistanzen zwischen den Fremdurteilern (FB) zu einer Person sind signifikant kleiner als die Interclusterdistanzen zwischen Beurteilern zu verschiedenen Personen.

H1b: Beim Vergleich der Intraclusterdistanzen (d.h. Distanzen innerhalb einer beurteilten Person) zwischen den verschiedenen Fremdurteilern ergeben sich überzufällig mehr kongruente als inkongruente Fremdbilder-Paare.

Basierend aus den Erkenntnissen der Literatur kann davon ausgegangen werden, dass sich dieser Effekt beim Vergleich zwischen Selbst- und Fremdbild umkehrt, d.h. es werden systematisch mehr inkongruente als kongruente Beurteilerpaare beim Vergleich von Selbst- und Fremdbild erwartet. Die zweite Hypothese lautet folglich:

H2: Beim Vergleich der Intraclusterdistanzen zwischen den verschiedenen Fremdurteilern und Selbsturteilen ergeben sich überzufällig mehr inkongruente als kongruente Selbstbild-Fremdbildpaare. Dies bedeutet, dass es über die Gesamtstichprobe betrachtet weniger kongruente Fremd- und Selbstbilder zu einer Person als inkongruente Fremd- /Selbstbilder-Paare gibt.

## 8.2 Methodik

Für die vorliegende Studie wurde ein kompetenzbasiertes Forced-Choice-Messverfahren entwickelt, welches lediglich das Profil und nicht die Profilhöhe, d.h. den absoluten Ausprägungsgrad berufsrelevanter Kompetenzen, abfragt.[17] Das Verfahren sieht vor, die einzelnen Kompetenzbegriffe und deren Definitionen in einem anwenderfreundlichen Online-Tool in zwei Schritten in eine Rangreihe zu bringen. Durch dieses Verfahren werden Kompetenzprofile gebildet, welche die relativen Stärken und Schwächen einer Person abbildet. Führungskräfte aus dem mittleren und oberen Managements aus unterschiedlichen Fachbereichen wurden mittels dieses onlinebasierten Kompetenzerhebungsverfahrens befragt.

In einem ersten Schritt wurden die Führungskräfte gebeten, die 15 Kompetenzdimensionen in drei Kategorien zu teilen, je nach Selbsteinschätzung des Ausprägungsgrades, von „weniger stark ausgeprägt“ bis „sehr stark ausgeprägt“ (Vgl. Abbildung 4).

[17] Die hier vorliegende Studie wurden mit denselben Daten durchgeführt wie die im vorangehenden Forschungsbericht IV.

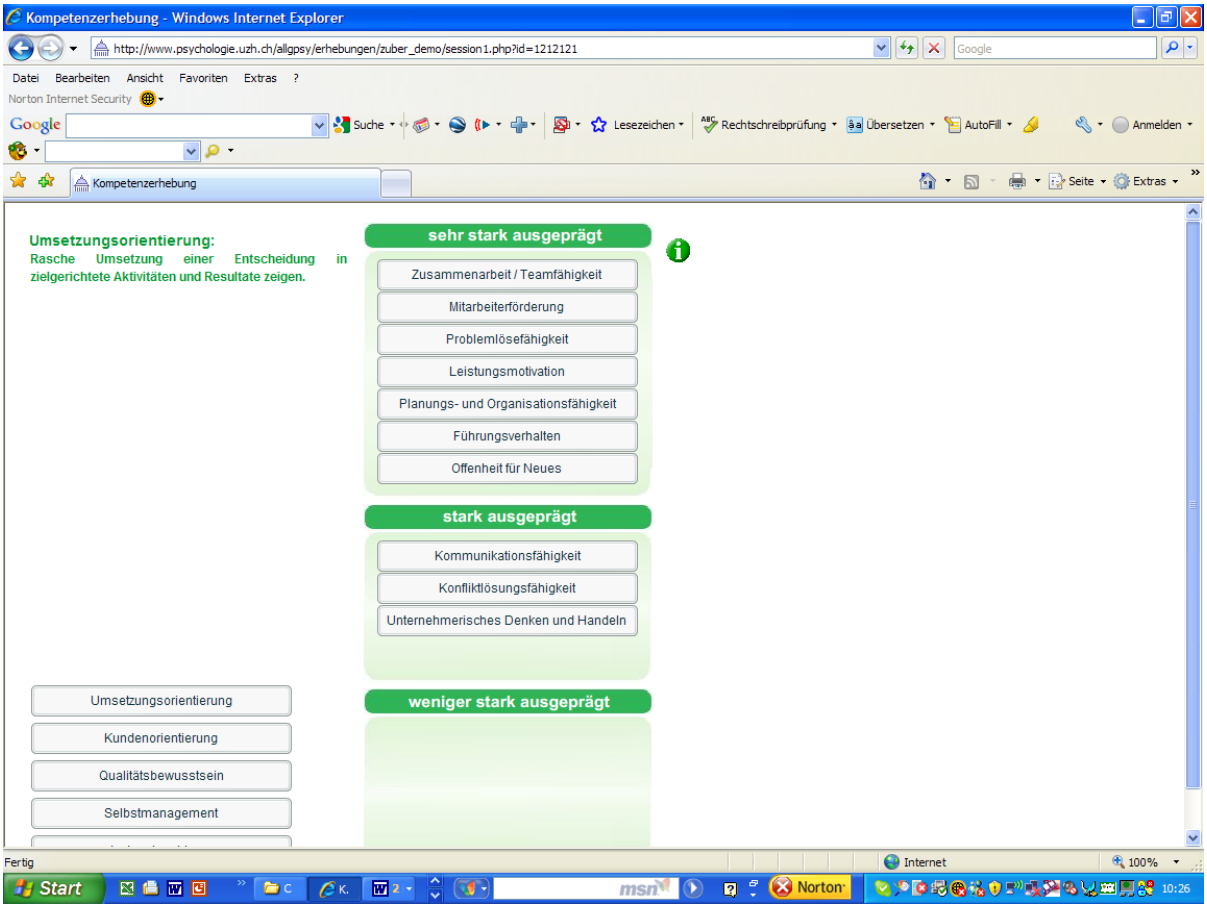



Abbildung 4: Kompetenzprofilbildung durch onlinebasiertes Forced-Choice-Tool, 1. Schritt

In einem zweiten Schritt (vgl. Abbildung 5) wurden die Versuchspersonen gebeten, die Kompetenzen pro Kategorie in eine Rangreihe zu bringen. Durch dieses Verfahren wurde pro Versuchsperson ein auf Rangplatz basierendes Kompetenzprofil erstellt. Somit wurde die Datenbasis für Rangkorrelationen zwischen Kompetenzprofilen geschaffen. Die Rangkorrelationen beruhen jeweils auf 15 Werten pro Person.

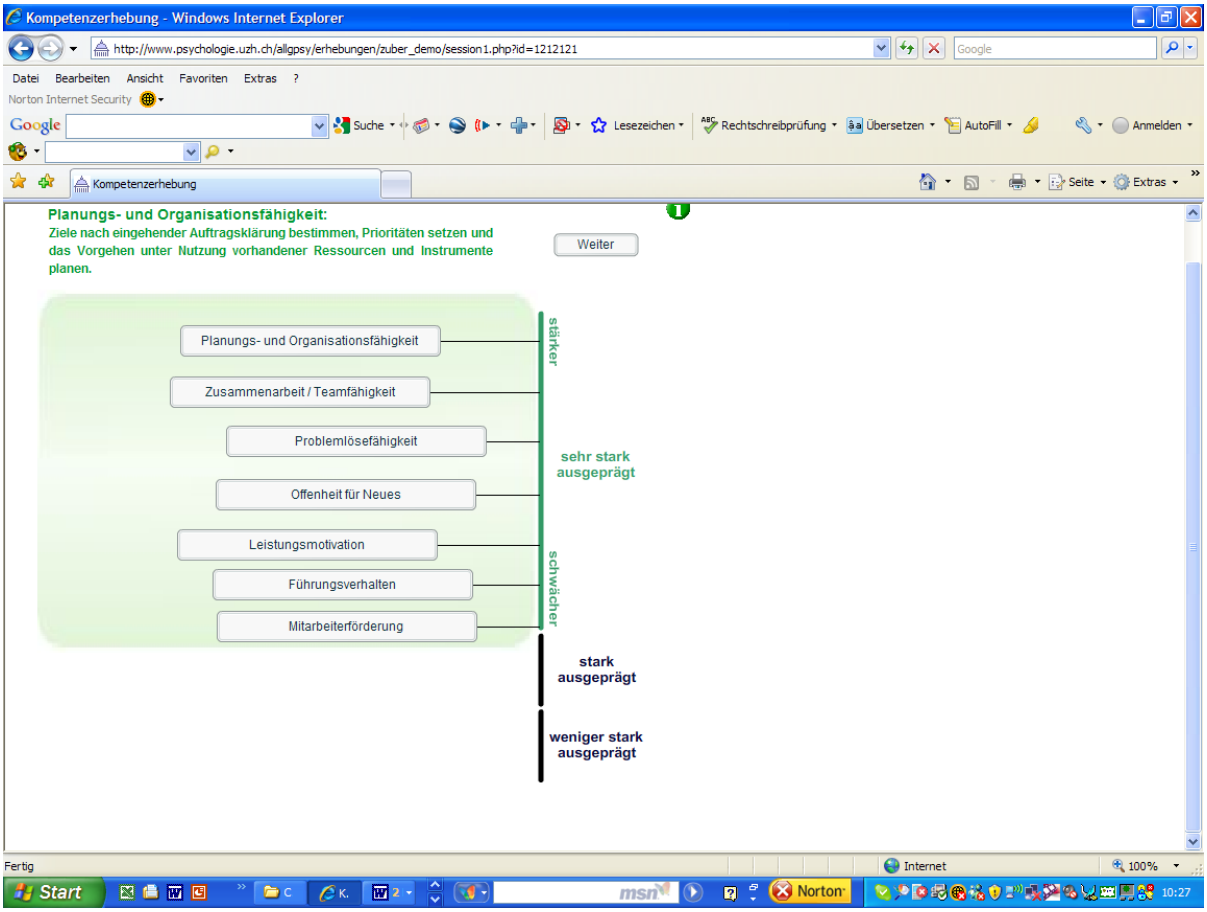



Abbildung 5: Kompetenzprofilbildung durch onlinebasiertes Forced-Choice-Tool, 2. Schritt

Basierend auf Selbst- und Fremdeinschätzungen zu zwei Erhebungszeitpunkten im Abstand von 2 Wochen wurden insgesamt 41 Kompetenzprofile erhoben, wobei 28 Versuchspersonen die Messbedingungen[18] erfüllten und somit in die Auswertung flossen.

Der Fokus der Verwendung der Daten liegt in dieser Studie jedoch nicht in der Visualisierung der Kompetenzprofile mittels Property Fitting, sondern in der Übereinstimmung verschiedener Beurteiler. Möchte man die Übereinstimmung zwischen Personen bzw. deren Sichtweisen anderer Personen messen, so muss man sich die Frage stellen, wie diese Übereinstimmung methodisch operationalisiert werden kann. Probleme in der methodischen Operationalisierung von „Übereinstimmung“ betreffen die Aggregation von Ratings sowie die Wahl eines geeigneten Diskrepanz-Masses. Dazu wurden die Profile aller Beurteiler über alle Beurteilte auf der Basis ihrer paarweisen Pearson-Korrelationen zunächst mittels NMDS in einen zweidimensionalen euklidischen Raum skaliert. Die Unähnlichkeit je zweier Profile

[18] Als Messbedingung wurden folgende drei Kriterien herangezogen: 1. Teilnahme an beiden Erhebungszeitpunkten, 2. Stabilität des Selbstbildes (Korrelation zwischen Selbstbildern zu Messzeitpunkt 1 und

wurde dann in diesem Raum als Distanz zwischen den beiden Punkten gemessen, die diese Profile repräsentieren.

Um die eingangs erwähnten Hypothesen zu überprüfen, wurden t-Tests für unabhängige Stichproben angewendet, um die Mittelwertsunterschiede der Intra- und Interclusterdistanzen zwischen den Fremdbeurteilern sowie zwischen Selbst- und Fremdbeurteilern zu überprüfen.

Wie in den Hypothesen ausgeführt wurde, wird nicht erwartet, dass alle Fremdbeurteiler eine einheitliche Sicht auf die einzelnen Beurteilten haben. Deswegen lautet die graduell formulierte Hypothese, dass zwischen den Fremdbeurteilern eine grössere Kongruenz als Inkongruenz in Bezug auf die jeweiligen Beurteilten besteht. Um zu beantworten, inwiefern diese Hypothese zutrifft bzw. inwieweit sich die Fremdbeurteiler untereinander einig sind, wurde ein Kongruenzmass herangezogen. Dabei wurde der Mittelwert aller extrahierten Distanzen aus der NDMS-Karte bestimmt und davon eine Standardabweichung aller Distanzen subtrahiert. Dieses Mass für die obere Grenze eine „kleinen“ Distanz zwischen zwei Objekten entspricht bei normalverteilten Daten circa einem Sechstel der Distanzen. Diejenigen Beurteilerpaare, welche unter diesem kritischen Kriterium (Mittelwert minus eine Standardabweichung aller Distanzen) liegen, wurden als „kongruente“ Paare definiert und so interpretiert, dass hier eine hinlänglich einheitliche Meinung über das Kompetenzprofil eines Beurteilten vorliegt. Diejenigen Paare, die darüber liegen, wurden als inkongruente Paare bezeichnet. Dieses Mass wurde sowohl beim Vergleich zwischen Fremdbeurteilern untereinander zu einer jeweiligen Person als auch zwischen Selbst- und Fremdurteilen angewendet.

## 8.3 Ergebnisse

Die Ergebnisse werden in drei Teilen berichtet: Zunächst erfolgt die Auswertung der Korrelationsmatrizen durch NMDS. In einem zweiten Teil wird die Übereinstimmung zwischen Fremdbildern zu jeweils derselben Person betrachtet. Dabei werden die Ergebnisse sowohl auf korrelativer Ebene als auch mittels Distanzmassen aus NDMS Karten präsentiert und miteinander verglichen. In einem dritten Teil wird dann die Übereinstimmung zwischen Fremdbildern und Selbstbildern berechnet und dargestellt.

Beim Vergleich zwischen der Kongruenz von Selbstbild und Fremdbild wurden nur diejenigen Beurteilerpaare berücksichtigt, bei denen das Selbstbild zwischen Messzeitpunkt 1 (t1) und Messzeitpunkt 2 (t2) stabil ist, d.h. die Distanz zwischen SB1 und SB2 unter dem

---

Messzeitpunkt von mind. r=.5), 3. Dauer der Erhebung: 3 Personen wurden aufgrund unrealistisch kurzer Erhebungsdauer (weniger als 1 Minute) aus der Untersuchung entfernt.

definierten Wert liegt. Eine Person (VP25) erfüllte dieses Kriterium nicht und wurde bei der Darstellung der Ergebnisse ausgeschlossen.

In der Folge werden die NDMS Karten aller Beurteiler und Beurteilten abgebildet. Die Karten zeigen die Clusterung bzw. die Übereinstimmung von Beurteilern zu einzelnen Beurteilten. Die Fremdurteile (FB) und Selbsturteile (SB) zu einer Person wurden jeweils in einer anderen Schattierung abgebildet. Im Dienste der Übersichtlichkeit wurden die Beurteilten zufällig in drei Gruppen aufgeteilt. Und zwar fand diese Zuordnung für 27 der 28 Beurteilten statt, so dass der Ausschluss von Person 25 sich auf die dritte Karte auswirkt.

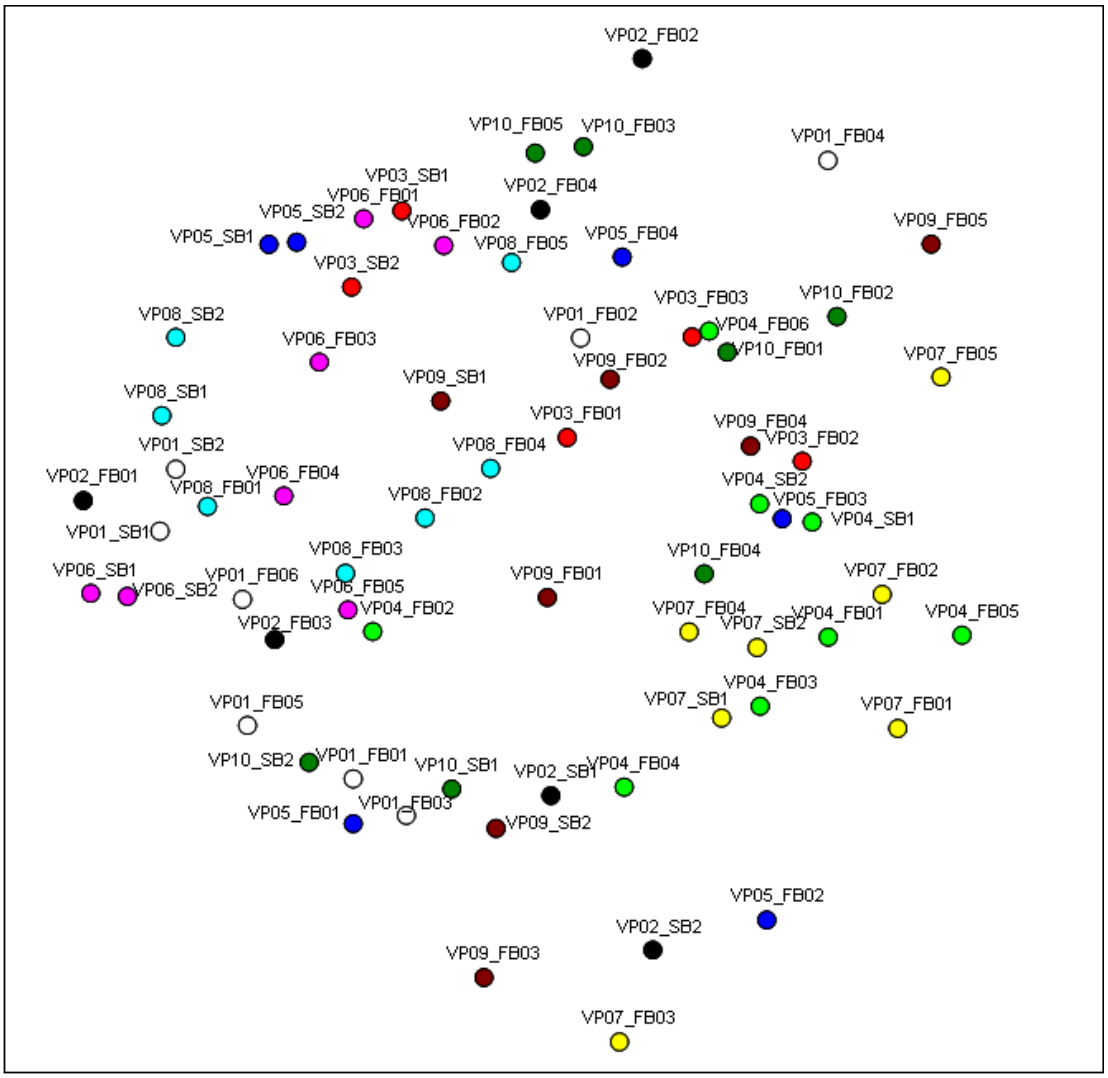



Abbildung 6: Gruppe 1- 10 Beurteilte (Fremd- und Selbstbild)

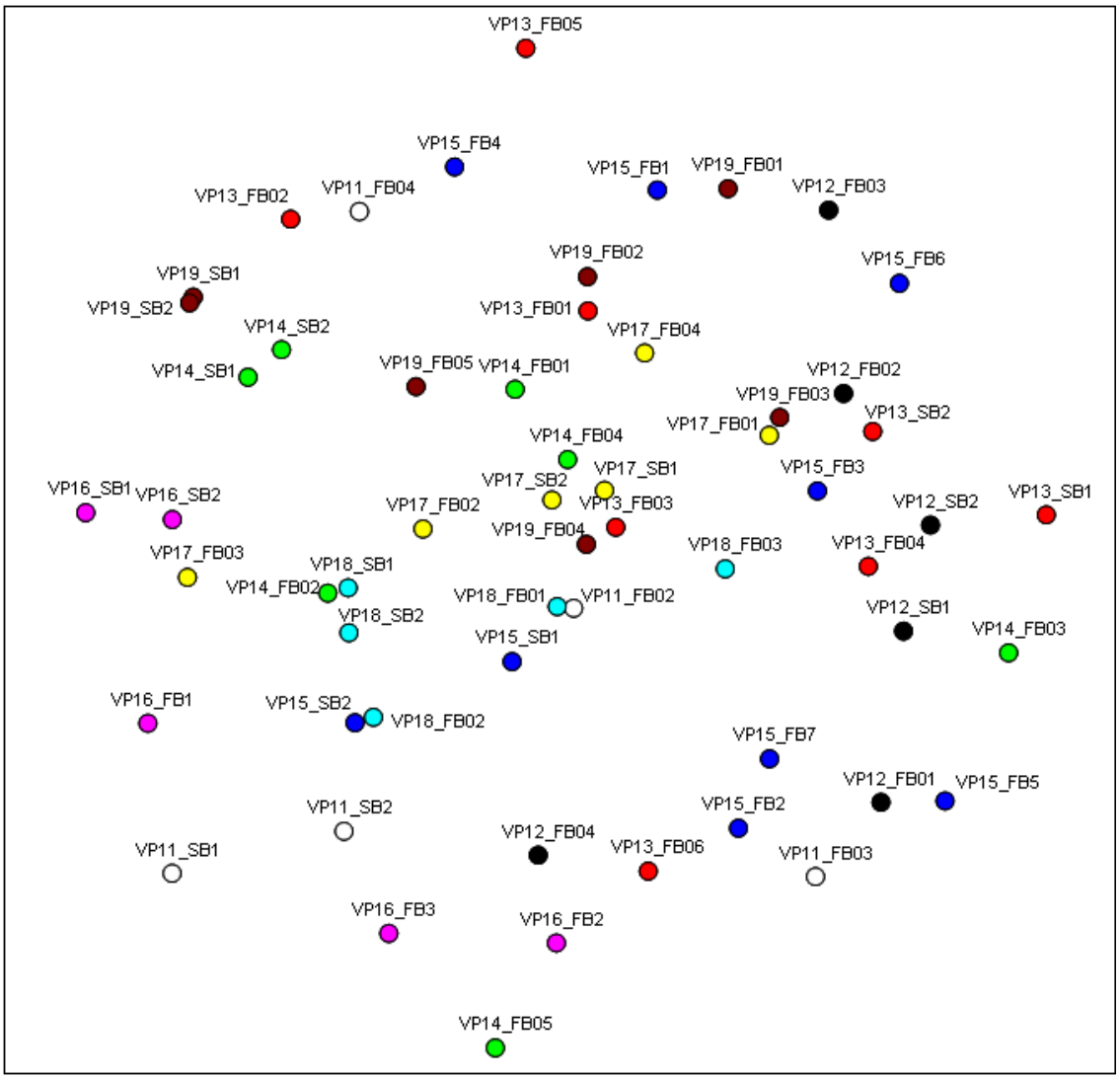



Abbildung 7: Gruppe 11- 19 Beurteilte (Fremd- und Selbstbild)

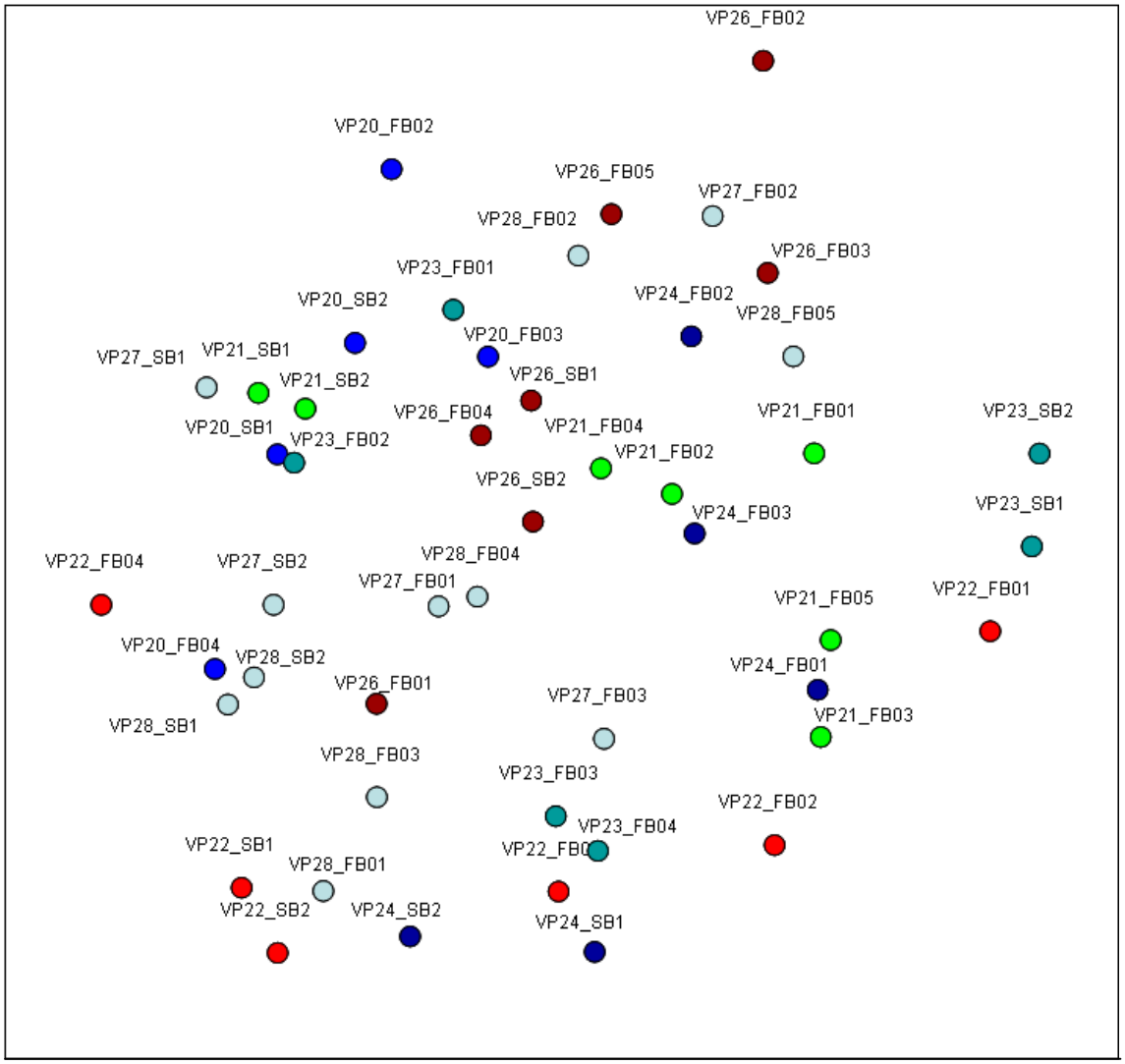



Abbildung 8: Gruppe 20- 28 Beurteilte (Fremd- und Selbstbild), ohne VP25

Auf den ersten Blick lässt sich eine gewisse Clusterung der Fremdurteile erkennen. Allerdings gibt es, wie in unserem Modell einer sozialen Realität vermutet, auch idiosynkratische Einzelsichten einzelner Beurteiler.

Betrachtet man die Daten auf korrelativer Ebene, so beträgt die mittlere Korrelation aller Fremdurteile (FB) zu jeweils identischen Personen $\bar{r}$ = 0.21. Die mittlere Standard-abweichung beträgt dabei $\bar{s}$ = 0.28. Diese Korrelation ist nicht sonderlich hoch. Vergleicht man sie jedoch mit der mittleren Korrelation aller Beurteilerpaare, $\bar{r}$ = 0.07 und $\bar{s}$ = 0.28, so lässt sich doch ein bestimmter übereinstimmender Effekt erkennen.

Um die Übereinstimmung zwischen den Fremdurteilern systematisch zu erfassen und diesen Effekt noch besser zu illustrieren, wurden aus der NMDS Karte die Distanzen zwischen den Punkten extrahiert, wodurch sich die Intra- und die Interclusterdistanzen der Fremdbeurteiler miteinander vergleichen lassen.

Die mittlere Distanz aller Intraclusterdistanzen der Fremdbeurteiler beträgt 1.09 bei einer Standardabweichung von 0.61. Die mittlere Distanz aller Interclusterdistanzen der Fremdbeurteiler über alle Personen beträgt 1.30, bei einer Standardabweichung von 0.60. Prüft man diesen Unterschied inferenzstatistisch mit t-Test für unabhängige Stichproben, so ergibt sich ein hochsignifikantes Resultat von $p < 0.001$ mit einer Effektstärke von $d=0.34$.

In dieser Studie interessiert gemäss dem zweiten Aspekt unserer ersten Hypothese (H1b) vor allem die Frage, ob bei verschiedenen Fremdbeurteilern zu einer Person vermehrt eine kongruente Sicht einer Person besteht oder ob vielmehr Einzelsichtweisen vorherrschen. Wie im methodischen Teil beschrieben, werden solche Beurteilerpaare zu einer Person als *kongruent* bezeichnet, wenn ihre Distanz eine Standardabweichung unter dem Mittel aller Distanzen liegt. Die in den ersten beiden Spalten der unten dargestellten Tabelle zeigen die Ergebnisse der Kongruenz bzw. Inkongruenz von Fremdbeurteilerpaaren zu jeweils einer Person:

| | **FB - FB kongruent** | **FB – FB inkongruent** | **SB - FB kongruent** | **SB - FB inkongruent** |
|---|---|---|---|---|
| **VP 1** | 4 | 2 | 1 | 5 |
| **VP 2** | 2 | 2 | 1 | 3 |
| **VP 3** | 3 | 0 | 0 | 3 |
| **VP 4** | 4 | 2 | 3 | 3 |
| **VP 5** | 0 | 4 | 0 | 4 |
| **VP 6** | 5 | 0 | 1 | 4 |
| **VP 7** | 2 | 3 | 3 | 2 |
| **VP 8** | 5 | 0 | 1 | 4 |
| **VP 9** | 2 | 3 | 0 | 5 |
| **VP 10** | 4 | 1 | 0 | 5 |
| **VP 11** | 2 | 2 | 1 | 3 |
| **VP 12** | 0 | 4 | 2 | 2 |
| **VP 13** | 5 | 1 | 2 | 4 |
| **VP 14** | 3 | 2 | 2 | 3 |
| **VP 15** | 7 | 0 | 2 | 5 |
| **VP 16** | 3 | 0 | 1 | 2 |
| **VP 17** | 3 | 1 | 2 | 2 |
| **VP 18** | 0 | 3 | 2 | 1 |
| **VP 19** | 5 | 0 | 0 | 5 |
| **VP 20** | 3 | 1 | 3 | 1 |
| **VP 21** | 4 | 1 | 1 | 4 |
| **VP 22** | 2 | 2 | 2 | 2 |
| **VP 23** | 4 | 0 | 0 | 4 |
| **VP 24** | 2 | 1 | 0 | 3 |
| **VP 26** | 2 | 3 | 2 | 3 |
| **VP 26** | 2 | 1 | 1 | 2 |
| **VP 28** | 3 | 2 | 1 | 4 |
| ∑ | **81** | **41** | **34** | **88** |

Tabelle 1: Vergleich FB – FB (Spalten 1 und 2) und Vergleich SB – FB (Spalten 3 und 4)

Die Zahlen der ersten zwei Spalten der Tabelle 1 zeigen, dass bei diesem angewendeten Kriterium zwei Drittel aller Distanzpaare von Fremdbeurteilern zu einer Person kongruent sind. Die Korrelationen der kongruenten Fremdbeurteilerpaare ergeben im Mittel $\bar{r}$ = 0.42, $\bar{s}$ = 0.16, die Korrelationen der inkongruenten Fremdbeurteilerpaare ergeben im Mittel $\bar{r}$ = 0.14, $\bar{s}$ = 0.18.

Im dritten Teil der Ergebnisse werden nun die Zahlen zwischen Selbst- und Fremdurteilen dargestellt (Vgl. Tabelle 1, Spalten 3 und 4). In Analogie zum Vergleich zwischen Fremdbeurteilern untereinander, wird auch beim Vergleich von Selbst- und Fremdurteil als Kongruenzmass der Mittelwert aller Distanzen minus eine Standardabweichung als *kritischen Wert* definiert. Da die Selbsturteile zu zwei Messzeitpunkten erhoben wurde, wurde der Mittelwert zwischen SB1 und SB2 berechnet.

Beim Vergleich zwischen der Kongruenz von Selbstbild und Fremdbild, wurden wie weiter oben erwähnt nur diejenigen Beurteilerpaare berücksichtigt, bei denen das Selbstbild zwischen Messpunkt 1 (t1) und Messpunkt 2 (t2) stabil ist, d.h. die Distanz zwischen SB1 und SB1 unter dem definierten kritischen Distanzwert liegt.

Stellt man die Zahlen der Fremdbildvergleiche (Spalten 1 und 2), den Zahlen der Selbst- / Fremdbildvergleiche gegenüber (Spalten 3 und 4), so stellt man fest, dass die Anzahl an kongruenten Paaren zwischen Selbst- und Fremdbildern stark reduziert ist. Es kann sogar eine Umkehrung des Effekts gemäss der eingangs formulierten Hypothese bestätigt werden. Wo beim Vergleich der Fremdbeurteiler untereinander zwei Drittel kongruente Paare ausgezählt werden konnten, können beim Vergleich zwischen Selbst- und Fremdbild bei Anwendung deselben *kritischen Kriteriums* nur noch ein Drittel kongruente Beurteilerpaare eruiert werden. Entsprechend ist auch die mittleren Korrelation zwischen allen Selbst- und Fremdurteilern mit $\bar{r}$ = .19, $\bar{s}$ =.16 entsprechend tiefer und unterscheidet sich kaum mehr von einer Null-Korrelation.

In Bezug auf unsere Hypothesen, können wir folgende Ergebnisse zusammenfassen.

Die Fremdbilder zu einer Person sind sich insgesamt ähnlicher als die Fremdbilder über alle Personen hinweg betrachtet. Dies konnte mit dem Vergleich von Intra- vs. Interclusterdistanzen von Fremdbildern gezeigt werden, wobei der Unterschied hochsignifikant wurde. Basierend auf den NMDS Karten konnte weiter gezeigt werden, dass nicht alle Fremdbeurteiler die gleiche Sicht einer Person haben, sich jedoch eine Mehrheit (zwei Drittel der Fremdbeurteiler-Paare erfüllen das kritische Kriterium) herauskristallisiert, die eine homogene bzw. kongruente Sichtweise einer Person haben. Unsere erste Hypothese (H1a und H1b) kann demnach vollumfänglich angenommen werden.

Auch unsere zweite Hypothese konnte mit den hier dargestellten Ergebnissen bestätigt werden. Beim Vergleich zwischen Selbst- und Fremdbild kippt der Effekt und die Diskrepanz zwischen Selbst- und Fremdbilder fällt relativ markant aus. Die Selbsturteile stimmen im Mittel nicht mit den Fremdurteilen überein. Welche Implikationen diese Erkenntnisse für die Personalbeurteilung haben und welche weiteren Forschungsanstrengungen unternommen werden müssen, wird im abschliessenden Kapitel erörtert.

## 8.4 Diskussion

Ziel der vorliegenden Studie war es, die Befundlage zur Übereinstimmung von Beurteilungsdaten im Bereich der multiperspektivischen Führungsbeurteilung anhand eines Verfahrens zu untersuchen, welches sich voll und ganz auf die Kompetenzprofile konzentriert und die allgemeine Profilhöhe eines Beurteilten ausser Acht lässt. Die Ergebnisse haben gezeigt, dass auch bei ipsativer Messung mit einem Forced-Choice-Format keine höhere Übereinstimmung zwischen Selbst- und Fremdurteil zu finden ist. Wie bei den gängigen Fragebogen-Formaten bewegen sich die Korrelationen bei ipsativer Messung auch um den Wert r = 0.20. Dieses Ergebnis liefert einen Beitrag zur Frage, ob die tiefe Übereinstimmung auf das Messinstrument zurückzuführen ist. Ohne diese Frage abschliessend beantworten zu können, so deutet dieses Ergebnis doch darauf hin, dass die geringe Übereinstimmung nicht im Wesentlichen auf das Messinstrument zurückzuführen ist. Auch mit einem Forced-Choice-Verfahren scheint sich die Inkommensurabilität zwischen Selbst- und Fremdbild zu bestätigen. Dies kann daher rühren, dass die Fremdbeurteiler aus sehr unterschiedlichen Perspektiven urteilen bzw. nicht in gleichem Umfang in Kontakt mit dem Beurteilten waren. Leider wurde aus Datenschutzgründen die Erhebung der Fremdbeurteilung vollkommen anonymisiert, so dass keine Rückschlüsse auf die berufliche Beziehung zwischen Beurteiler und Beurteilten gezogen werden können, was die Interpretation der geringen Korrelationen einzelner Personen schwierig macht. Dies zu tun war jedoch auch nicht das vordergründige Ziel der Studie. Vielmehr lag das Interesse dieser Studie in der Visualisierung der Übereinstimmung mittels NMDS, welche eine differenziertere Betrachtung der Daten ermöglicht als die blosse Interpretation von Korrelationskoeffizienten. Hier konnte ein wesentlicher Beitrag geleistet werden, indem mittels euklidischen Karten ein Modell der sozialen Realität von Beurteilungsdaten dargestellt werden konnte. Dabei wurde ersichtlich, dass es bestimmte Persönlichkeitstypen, bzw. Führungskräfte mit einem bestimmten Kompetenzprofil gibt, die aufgrund ihrer Profilähnlichkeit nahe beieinander in der Karte abgebildet werden. Eine Führungskraft wird dabei natürlich nicht von allen Personen gleich

betrachtet, sondern es gibt bei der Einschätzung eine gewisse Streuung. Dennoch ist die unterschiedliche Wahrnehmung des Kompetenzprofils nicht zufällig, sondern clustert im Raum in mehr oder weniger homogene Gruppen, die mehrere Personen mit ähnlicher Ausprägung von Kompetenzen umfassen. Die Überprüfung der Homogenität oder anders ausgedrückt der Beurteilerübereinstimmung wurde durch die Inter- und Intraclusterdistanzen aus den NMDS Karten ermittelt. Dabei hat sich gezeigt, dass die Intracluster-Distanzen signifikant kleiner sind als die Interclusterdistanzen über alle Beurteilten. Dies bedeutet, dass Gemeinsamkeiten in den Kompetenzprofilen zu jedem Beurteilten tatsächlich vorhanden sind und von einer Mehrheit von Beurteilern auch so erkannt werden.

Durch den Vergleich der Kongruenz einzelner Beurteilerpaare zu ein und derselben Person konnte weiter gezeigt werden, dass die Mehrheit der Beurteiler ein ähnliches Bild dieser Person hat (81 kongruente Paare, 41 inkongruente Paare). Wie in den Ergebnissen berichtet, kippt dieser Effekt, wenn man die Anzahl kongruenter Beurteilerpaare zwischen Selbst- und Fremdbild vergleicht. Hier konnten nur 38 kongruente, dafür aber 88 inkongruente Paare eruiert werden.

Neben der methodischen Betrachtung der Ergebnisse, sollen an dieser Stelle einige Überlegungen zur inhaltlichen Interpretation der Daten gemacht werden und mögliche Erklärungsversuche zur geringen Übereinstimmung bestimmter Beurteiler gefunden werden.

Menschliche Beurteilung ist nie eine "1:1-Abbildung" einer Wirklichkeit, sondern immer eine individuelle Beschreibung, die auch genährt ist von persönlichen Interessen, Motiven, Erwartungen und mentalen Verarbeitungsprozessen der Rater. Auf unsere Thematik angewandt bedeutet dies, dass sich eine multiperspektivische Beschreibung aus so vielen Wirklichkeiten wie Beurteilern zusammensetzen muss und deshalb keine totale Übereinstimmung anvisiert werden kann. Doch warum differieren diese Wirklichkeiten so stark? Und warum unterscheiden sich die Wirklichkeiten der sich selbst beurteilenden Führungskräfte so stark von denen anderer Beurteiler?

Vieles deutet darauf hin, dass ein Zusammenspiel verschiedener Ursachen für die Ergebnisse verantwortlich ist. In dieser abschliessenden Diskussion steht entsprechend des aktuellen Erkenntnisstandes die Selbst-Fremd-Übereinstimmung im Mittelpunkt. Einzelne Erklärungsansätze werden auf die Fremd-Fremd-Übereinstimmung übertragen. Auf verschiedenen Ebenen können Faktoren die Verzerrung von Selbst- oder Fremdbild beeinflussen. Bei deren Darstellung wird zurückgegriffen auf allgemeine Beschreibungen des Beurteilungsprozesses, dessen Ergebnis sich aus drei Klassen von Faktoren ergibt (nach Wherry & Bartlett, 1982; erweitertes Begriffsverständnis bei Scullen et al., 2000):

erstens aus dem Leistungsverhalten der Fokusperson (Ratee-Leistung) selbst, zweitens aus verschiedenen Prozessen der Informationsverarbeitung im weitesten Sinne, die der Aussage des Raters (Selbst- oder Fremdrater) zugrunde liegen (Rater-Bias) sowie drittens aus dem Messfehler. Die Ratee-Leistung ergibt sich aus dem generellen Leistungsniveau der Fokusperson sowie ihren speziellen Leistungen auf einzelnen Dimensionen.[19] Idealerweise geben Leistungsratings nur dieses aktuelle Leistungsverhalten der Fokusperson wieder.

Doch die Korrelationen zwischen Leistungsratings und objektiven Leistungsmassen sind nach meta-analytischen Schätzungen allenfalls mittelmässig (maximal r = .39; Bommer et al., 1995; Conway et al., 2001), da Ratingantworten kontaminiert und defizient sind. Während Bommer et al. (1995) Abwärtsratings als subjektive Masse heranzogen, waren es bei Conway et al. (2001) Aufwärts- und Peerratings. Ausserdem gibt es keinen Hinweis auf deutlich andere Zusammenhänge in der Beurteilung von Führungsverhalten, auch wenn die zitierten Meta-Analysen berufliche Leistung im Allgemeinen betrachten. Diese Annahme wird durch die Befunde zum Zusammenhang zwischen Aufwärtsratings und Erfolgskriterien unterstützt. Der nur mittelmässige Zusammenhang zwischen Ratings und objektiven Kriterien darf aber nicht als Limit, das nicht übertroffen werden kann, bewertet werden. Es zeigt sich nämlich, dass subjektive Ratings und objektive Masse sehr stark korrelieren (r = .71), wenn sie so konstruiert sind, dass sie genau die gleiche Leistungsdimension erfassen (Meta-Analyse von Conway et al., 2001). Zu bedenken gilt weiter, dass auch objektive Masse defizient sind, also auch diese zu den geringen Korrelationen beitragen (vgl. Murphy, Cleveland & Mohler, 2001). So erfasst beispielsweise das objektive Mass der Fehltage eher die Führungsleistung, die sich im Bereich der Motivierung und Mitarbeiterorientierung abspielt, vernachlässigt aber die Führungsleistung, die sich mit der Aussendarstellung des Unternehmens befasst. Für komplexe Bereiche wie das Führungsverhalten ist es schwierig, ein einzelnes relevantes, objektives Kriterium zu finden. Wenn aber Ratings mit objektiven Massen in aller Regel nur gering korrelieren, die Ratee-Leistung also nur ansatzweise abgebildet wird, dann bedeutet das, dass ein Teil der Ursachen für die mangelnde Übereinstimmung von Selbst- und Fremdbild auf der Raterseite, am Rater-Bias, liegt: Zum einen kann sie in den unterschiedlichen organisationalen Perspektiven von Selbst- und Fremdratern begründet sein (perspektivenbezogener Bias) - Dies liegt deshalb sehr nahe, weil die Übereinstimmung von Ratern aus derselben Perspektive höher ist als von Ratern aus verschiedenen Perspektiven (Borman, 1974). Zum anderen kann eine mangelnde Selbst-Fremd-Übereinstimmung aber

[19] Da manche Leistungsvoraussetzungen für verschiedene Leistungsfacetten sehr ähnlich sind, ist von einem „wahren Halo“ auszugehen, der folglich nicht als Fehler missinterpretiert werden darf.

auch dadurch zustande kommen, dass die Beurteiler das zu beurteilende Führungsverhalten aufgrund individueller Faktoren verzerren. Dies stellt zugegebenermassen eine Kategorie mit nur sehr vagen Grenzen dar: "Consequently, we use the term idiosyncratic rater effects to include all of the effects associated with individual raters." (Mount & Scullen, 2001, p. 165). Diese Kategorie umfasst also – im Gegensatz zum unsystematischen Fehleranteil – die systematische Varianz, die mit einzelnen Ratern, nicht aber mit dem Ratee-Verhalten oder den Ratings anderer verbunden ist (nach Scullen, Mount & Goff, 2000). Darunter fallen individuelle Beurteilertendenzen, die beispielsweise auf dem Motiv, sich positiv darzustellen, gründen. Bei ipsativer Messung würde man in dieser Studie nicht von „positiv“ sprechen, sondern eher von erwartungskonform im Hinblick auf das Anforderungsprofil. Bei ipsativer Messung kann demnach die tiefe Korrelation nicht durch die Selbstüberhöhungstendenz (Above-average Effekt) erklärt werden.

Eine weitere Erklärungskomponente könnte die mangelnde Beobachtbarkeit und Transparenz von Entscheidungsprozessen bei den Ratern sein (Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988; Heidemeier & Moser, 2002). Wenn die zu beurteilenden Verhaltensweisen mit Leichtigkeit beobachtet werden können, dann steigt die Güte der Ratings (Ratingtheorie von Wherry & Bartlett, 1982). Doch die Beobachtung von Führung und anderer fachübergreifenden Kompetenzen ist für Fremdbeurteiler häufig alles andere als leicht. Deshalb wird mangelnde Gelegenheit zur Beobachtung des zu beurteilenden Verhaltens seit langem als Faktor diskutiert, der die Reliabilität von Ratings mindert (Dunnette, 1966; Nagle, 1953). Empirische Bestätigung findet dieser Punkt erstmals bei Rothstein (1990), die einen starken Zusammenhang zwischen der Interrater-Reliabilität und der Dauer der Führungstätigkeit feststellte. Möglicherweise liegt dies auch daran, dass weniger Beobachtungsmöglichkeiten zu systematischen Beurteilungsfehlern führen (Rothstein, 1990).

Von Seiten der Selbsteinschätzung kann weiter argumentiert werden, dass Manager bei ihrer Selbstbeurteilung insgesamt stärker auf spezielles als auf globales Leistungsverhalten zurück greifen (Scullen, Mount & Goff, 2000, siehe S. 178). Da jedes Selbstkonzept differenzierter ist als ein Fremdbild, können Führungskräfte ihre Selbstbeurteilung auf spezielleren Aspekten gründen als Fremdbeurteiler. Zur Beurteilung mancher Führungsdimensionen, zum Beispiel Kommunikationsverhalten, Kritikfähigkeit oder Durchsetzungsvermögen, können zudem arbeits-irrelevante Selbstkonzepte (multiple Selbstkonzepte, Markus & Wurf, 1987) relevant werden. So ist die Selbstbeschreibung aber möglicherweise von einem allgemeinen Selbstbild gestützt, während es den Fremdbeobachtern besser gelingt, sich lediglich auf arbeitsbezogene

Situationen zu beziehen. Wenn jedoch auf unterschiedliche Information zurückgegriffen wird, korrelieren Selbst- und Fremdbild in der Folge kaum. Durch die Aktivierung mehrerer Selbstkonzepte steht dem Selbstbeurteiler auch ein grösserer Pool an Informationen zur Verfügung, woraus er – im Sinne der Selbstwerttheorie – zunächst diejenigen erinnert und abruft, die positiv sind. Und: Es wird nicht jede relevante Information zur Urteilsfindung bzw. zur Beantwortung von Items herangezogen. Indes geht es vielmehr um einen Suchprozess, der vorzeitig abgebrochen wird, sobald ausreichend Information zur Itembeantwortung verfügbar ist (vgl. Bodenhausen & Wyer, 1987).

Auch wenn die in der Literatur breit diskutierten Erklärungsansätze durchaus plausibel sind, so gehen wir noch einen Schritt weiter und postulieren eine gewisse Erkenntnisbarriere zwischen Selbst- und Fremdeinschätzung. In Anlehnung an die Wahrnehmungspsychologie könnte man auch von inkommensurablen Erfahrungswelten sprechen. Die Selbsteinschätzung fühlt sich kategorial anders an und ist psychologisch gesehen, ein zur Fremdeinschätzung völlig unterschiedlicher Beurteilungsprozess. Während es sich bei der Selbsteinschätzung um das persönliche Erleben handelt, also die Innenwelt einer Person betrifft, geht es bei der Fremdeinschätzung um die Wahrnehmung der Aussenwelt, in diesem Zusammenhang um die Wahrnehmung der Wirkung eines bestimmten Verhaltens.

In Übereinstimmung mit anderen Autoren (z.B. Becker et al., 2002; Borman, 1997; Moser, 1999; Murphy & Cleveland, 1995) muss eine unbefriedigende Forschungslage festgehalten werden, was die Erklärung der geringen Selbst-Fremd-Übereinstimmung von Leistungs- und Kompetenzratings angeht. Trotz der dünnen und uneindeutigen Ergebnislage bestehen für viele Erklärungen unterstützende Argumente. Obwohl sie teilweise schon sehr lange diskutiert werden, wurden sie jedoch selten gezielt untersucht.

Diese Studie hatte nicht zum Ziel, alle möglichen Erklärungsversuche für die geringe Übereinstimmung zwischen Selbst- und Fremdurteilen darzulegen. Vielmehr lag der Fokus darin zu zeigen, dass die mangelnde Übereinstimmung nicht auf das Fragebogenformat und die damit verbundenen Urteilstendenzen zurückzuführen ist. Aus den Ergebnissen geht hervor, dass die geringe Übereinstimmung zwischen Selbst- und Fremdurteil auch bei einem ganz anders strukturierten Messinstrument gefunden wird: Wie bei den Fragebogenformaten führt auch ein Forced-Choice-Verfahren zu höheren Beurteilungsdiskrepanzen zwischen Selbst- und Fremdurteil als zwischen Fremdbeurteilern untereinander.

Von noch grösserer Bedeutung als dieser Befund ist die methodische Herangehensweise, wie diese Diskrepanzen zwischen verschiedenen Beurteilern mittels NMDS-Karten auf an-

schauliche Weise dargestellt wurden. Hier ergibt sich auch ein Bezug zur Praxis und zum in der Literatur häufig zitierten Konfliktpunkt, den Tornow (1993) einbrachte, indem er das wissenschaftliche und praktische Interesse an multiperspektivischen Beurteilungen nahezu unvereinbar nebeneinander stellte. Für die Forschung stehe die Messung an sich Zentrum des Interesses (z.B. die Ratinggenauigkeit), so dass eine Reduzierung der mangelnden Übereinstimmung zwischen den Ratern nur deshalb angestrebt würde, um den Ratingfehler zu minimieren. Die Praktiker hingegen wollten die Methode nutzen, um die individuelle und organisationale Effektivität zu steigern und um auf Lernprozesse zu setzen, die durch die Beurteilungen angestossen würden. Die Messung sei dann also nur Mittel zum Zweck. Der Lernprozess ist jedoch nur dann erfolgreich, wenn die Beurteilten die unterschiedlichen Sichtweisen auch nachvollziehen können. Die NMDS Karten ermöglichen gegenüber den üblichen Feedbackreports, welche die Ausprägungen verschiedener Kompetenzen mittels Balken- oder Spinnendiagrammen oft nur als Einzelsichten sinnvoll darstellen können, eine multiperspektivische Darstellung. So können in den NMDS Karten die verschiedenen Sichtweisen aller Beurteiler sowie Beurteilten auf einen Blick erfasst werden, was zum leichteren Erkennen der unterschiedlichen Sichtweisen beiträgt. Die NDMS Karten bieten damit eine anschauliche Diskussionsgrundlage für Mitarbeiter und Führungskräfte, um die unterschiedliche Selbst- und Fremdsicht aus multiplen Perspektiven zu beleuchten. Anstatt nach mehr Ursachen für die Diskrepanzen multiperspektivischer Beurteilung zu forschen, wäre es vielleicht an der Zeit, die sozialwissenschaftliche Realität, d.h. die Subjektivität von Beurteilungen zu akzeptieren und nach Methoden zu forschen, wie mit dieser Unterschiedlichkeit sinnvoll umgegangen werden kann.

## 8.5 Literatur


Atwater, L.,Ostroff, Ch., Yammarino, F. & Fleenor, J. (1998). Self-other agreement:Does it really matter? *Personnel Psychology*, 51, 577-598.

Atwater, L., Roush,P. & Fischthal, A. (1995). The influence of upward feedback on self- andfollower ratingsof leadership. *Personnel Psychology, 48*, 35-59.

Atwater, L. &Yammarino, F. (1992). Does self-other agreement on leadership perceptions moderate thevalidity of leadership and performance predictions? *Personnel Psychology, 45*, 141-164.

Bailey, R. C. &Bailey K. G. (1974). Self-perceptions of scholastic ability at four gradelevels. *Journal of Genetic Psychology, 124*, 197-212.

Baril, G. L.,Ayman, R. & Palmiter, D. J. (1994). Measuring leader behavior: Moderatorsof discrepant selfand subordinate descriptions. *Journal of Applied Social Psychology, 24*, 82-94.

Bass, B. M. (1985).*Leadership and performance beyond expectations.* New York: Free Press.

Bass, M. B. &Yammarino, F. J. (1991). Congruence of self and others' leadership ratings of naval officers for understanding successful performance. *Applied Psychology: An International Review, 40*, 437-454.

Becker, J., Ayman,R., & Korabik, K. (2002). Discrepancies in self/subordinates' perception of leader behavior.*Group & Organization Management*, 27(2), 226-244.

Bodenhausen,G.V./ Wyer, R.S. (1987): Social cognition and social reality: Information acquisitionand use in the laboratory and the real world. In: Hippler, H.-J./ Schwarz, N./Sudman, S. (Hrsg.) *Social information processing and survey methodology*, New York:Springer, S.6-41.

Bommer, W. H.,Johnson, L. J., Rich, G. A, Podsakoff, P. M. & Mackenzie, S. B. (1995). On the interchange ability of objective and subjective measures of employee performance: a meta-analysis.*Personnel Psychology, 48,* 587- 605.

Borman, W. C.(1974). The rating of individuals in organizations: An alternative approach. *Organizational Behavior and Human Performance*, 12, 105−124.

Borman, W. C.(1997). 360 ratings: An analysis of assumptions and research agenda for evaluating their validity. *Human Resource Management Review*, 7, 299−315.

Bortz, J. (2005): Statistik für Human- und Sozialwissenschaftler. 6. Auflage, Springer, Berlin.

Brutus, S. &Derayeh, M. (2002). Multisource assessment programs in organizations: An insider's perspective.*Human Resource Development Quarterly, 13,* 187-203.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E., (2005). Reconsidering Forced-Choice Item Formats for Applicant Personality Assessment. *Human Performance*, 18, 3, 267-307.

Church, A. H.(1997b). Do you see what I see? An exploration of congruence in ratings from multiple perspectives.*Journal of Applied Social Psychology, 27,* 983-1020.

Church, A. H.(1997a). Managerial self-awareness in high-performing individuals in organizations. *Journal of Applied Psychology, 83,* 281-292.

Conway, J. M. &Huffcutt, A. I. (1997). Psychometric properties of multisource performace ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance,10,* 331-360.

Conway, J. M.,Lombardo, K. & Sanders, K. C. (2001). A meta-analysis of Incremental Validity and Nomological Networks for Subordinate and Peer Ratings. *Human Performance, 14,* 267-303.

Drees, H. B. (1994). Untersuchungzur Validität eines Assessment Centers. Hinweise zur empirischen Überprüfung eines Assessment Centers unter besonderer Berücksichtigung unterschiedlicher Beobachtertrainings. Aachen: Dissertationsschrift.

Dunnette, M. D. (1966). Personnel selectionand placement. Belmont, CA: Wadsworth, 1966.

Edwards, M. R.& Ewen, A. J. (2000). 360°-Beurteilung: klareres Feedback, höhere Motivation und mehr Erfolg für alle Mitarbeiter. München: Beck Wirtschaftsverlag.

Fennekels,G. P. (2000). *Qualitative Führungsstilanalyse* (2. Aufl.). Göttingen: Hogrefe.

Fleenor, J. W.,McCauley, C. D. & Brutus, S. (1996). Self-other rating agreement and leader effectiveness. *Leadership Quarterly,* 7, 487-506.

Fittkau-Garthe, H.& Fittkau, B. (1971). Fragebogen zur Vorgesetzten-Verhaltens-Beschreibung (FVVB). Göttingen: Hogrefe.

Furnham, A. &Stringfield, P. (1998). Congruence in job-performance ratings: A study of 360°feedbackexamining self, manager, peers, and consultant ratings. *Human Relations, 51,* 517–538.

Harris, M. M. &Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisorratings. *Personnel Psychology, 41,* 43-62.

Hedge, J. W.,Borman, W. C. & Birkeland, S. A. (2001). History and development of multisource feedbackas a methodology. In D. W. Bracken, C. W. Timmreck & A. H. Church (Eds.), *The handbookof multisource feedback* (pp. 15-32). San Francisco: Jossey-Bass.

Heidemeier, H.& Moser, K. (2002). *Self-appraisal of job-performance*. Poster beim 43. Kongress derDeutschen Gesellschaft für Psychologie.

Heidemeier, H. & Moser, K. (2009) „Self-other agreement in job performance ratings: A meta-analytical test of a process model“, *Journal of Applied Psychology*, 94, 353-370.

Johns, G. (1981). Difference scoremeasures of organizational behavior variables: A critique. *Organizational Behavior and Human Performance,27,* 443-463.

Klimoski, R. J.& London, M. (1974). Role of the rater in performance appraisal. *Journal of Applied Psychology,59,* 445-451.

Koch, S. (2001). EigenverantwortlichesHandeln von Führungskräften. Schriftenreihe *Organisation &Personal* (Bd.10). München, Mering: Rainer Hampp.

Liebel, H. J. & Oechsler, W.A. (1994). *Handbuch Human-Resource-Management*. Wiesbaden: Gabler.

Mabe, P. A. & West, S. G.(1982). Validityof self evaluations of ability: A review and metaanalysis. *Journalof Applied Psychology, 67,* 280-296.

Malloy, T. E.,Yarlas, A., Montvilo, R. K. & Sugarman, D. B. (1996). Agreement and accuracy in children's interpersonal perceptions: A social relations analysis. *Journal of Personality and Social Psychology, 71,* 692-702.

Markus, H., &Wurf, E. (1987). The dynamic self-concept: A social psychological perspective.In M. R.Rosenzweig & L.W. Porter (Eds.), *Annual Review of Psychology,* 38, 299-337.

Meindl, J. R.(1993). Reinventing leadership: A radical, social psychological approach. In J.K. Murnighan(Ed.), *Social psychology in organizations: advances in theory and research*. EnglewoodCliffs: Prentice Hall.

Meindl, J. R.(1998). The romance of leadership as a follower-centric theory: a social constructionist approach. In Dansereau, F. & Yammarino, F. J. (Eds.), *Leader-ship: The multiple-level approaches.* Stamford: JAI Press.

Moser, K. (1999). Selbstbeurteilung beruflicher Leistung:Überblick und offene Fragen. P*sychologischeRundschau,* 50, 14-25.

Mount, M. K. &Scullen, S. E. (2001). Multisource feedback ratings: What do they really measure? In M. London (Ed.), *How people evaluate others in organizations* (S. 155-176). Mahwah:Lawrence Erlbaum Associates.

Murphy, K.R.,Cleveland, J.N. & Mohler, C.J. (2001). Reliability, validity, and meaningfulness of multisource ratings. In D. Bracken, C. Timmreck, and A. Church (Eds.), *Handbook of multisource feedback* (130-148). San Francisco: Jossey Bass.

Nagle, G . F. (1953).Criterion development. *Personnel Psychology*, 6, 271-289.

Nilsen, D. &Campbell, D. P. (1993). Self-observer rating discrepancies: Once an overrater, always an overrater? *Human Resource Management, 32,* 265-281.

Okazaki, S. (2002).Self-other agreement on affective distress scales in Asian Americans and White Americans. *Journal of Counseling Psychology, 49,* 428-437.

Parker, J. W.,Taylor, E. K., Barrett, R. S. & Martens, L. (1959). Rating scale content: Relationship between supervisory- and self-ratings. *Personnel Psychology, 12,* 49-63.

Rothstein, H. R.(1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, 75, 322-327.

Schuler, H. (2003).Beurteilung und Förderung beruflicher Leistung. 2. überarbeitete und erweiterte Auflage.Hogrefe, Göttingen.

Scullen, S. E.,Mount, M. K. & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85,* 956-970.

Smither, J. W.,London, M., Vasilopoulos, N. L., Reilly, R. R., Millsap, R. E. & Salvemini, N. (1995). An examination of the effects of an upward feedback program over time. *Personnel Psychology,48,* 1-34.

Stempfle, J., Hagmayer, Y.,Hübner, O., Iwanoff, Ch. & Kaufmann, St. (2004). Weiterentwicklung und Dynamisierung des Assessment Centers auf der Basis eines komplexen computersimuliertenProblemlöseszenarios: Das Challenge-AC der Thomas Cook AG. *Wirtschaftspsychologie, 2,* 23-36.

Sudman, S. Bradburn, N. M. &Schwarz, N. (1996). Thinking about answers. San Francisco: Jossey-Bass Tornow, W. (1993). Perceptions or reality: Is multi-perspective measurement a means or an end? *Human Resource Management, 32,* 221-229.

Tversky, A. &Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,*1124-1131.

Van Velsor, E.,Taylor, S. & Leslie, J. B. (1993). An examination of the relationships amongs selfperception accuracy, self-awareness, gender, and leader effectiveness. *Human Resource Management*, 32, 249-263.

Viswesvaran, C.,Ones, D. S. & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557–574.

Watzlawick, P. (1976). Wie wirklich ist die Wirklichkeit – Wahn, Täuschung, Verstehen; R. Pieper & Co. Verlag, München.

Wherry, R. J., Sr.,& Bartlett, C. J. (1982). The control of bias in ratings: a theory ofrating. *Personnel Psychology,*35, 521-551.

Wohlers, A. J.& London, M. (1989). Ratings of managerial characteristics – Evaluation difficulty, coworker agreement and self-awareness. *Personnel Psychology, 42,* 235-261.

# 9 Reliability of competency scaling depending on the rating and item format

## 9.1 Introduction

Despite the optimism that motivated distortion does not represent a serious threat to personality tests used to aid organizational decision making (Barrick & Mount, 1996; Hogan, Hogan, & Roberts, 1996; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Ones & Viswesvaran, 1998; Ones, Viswesvaran, & Reiss, 1996), evidence continues to emerge that attempts to improve scores on self-report personality inventories may be a problem in applicant samples. For example, when surveyed confidentially a substantial proportion of applicants admit to intentionally misrepresenting themselves in self-reports collected as part of the hiring process (Donovan, Dwight, & Hurtz, 2003; Holtgraves, 2004; McDaniel, Douglas, & Snell, 1997). Applicant scores on personality tests have also been shown to be inflated compared to non-applicants and to correlate more highly with measures of socially desirable responding (Hough, 1998; Rosse, Stecher, Miller, & Levin, 1998). Although the effects of applicant distortion continue to be debated in academic circles (Dilchert, Ones, Viswesvaran & Deller, 2006), substantial scepticism exists in industry regarding the use of self-report measures to facilitate hiring decisions. Concern that motivated applicants can easily distort personality measures remains the most widespread criticism organizational decision makers have of personality testing (Cook, 1993; Hogan & Hogan, 1992; Hogan et al., 1996). Disregarding such admonitions, many personality inventories popular in industry have been developed without concern over socially desirable responding due to the belief that self-enhancement represents valid trait variance and, therefore, is not a problem (see e.g., Costa & McCrae, 1992; Hogan & Hogan, 1992). By better addressing these concerns, applied psychologists may engender more confidence in the professionals our practice serves. Indeed, a recent survey of practitioners who work in the area of selection and assessment found that approximately 70% expressed preference for using a personality inventory that includes a method to deal with applicant distortion (Goffin & Christiansen, 2003). One of the earliest methodologies proposed for dealing with motivated distortion in personality assessment was to employ forced-choice item formats. These formats

result in response options equated in terms of perceived attractiveness so that respondents cannot simply describe themselves more favourably in an effort to create a positive impression (e.g., Berkshire, 1958; Edwards, 1959). In spite of some early promise (Zavala, 1965), the popularity of the forced-choice response format declined throughout the 1970s to the point that few professionals advocated their use. For example, in the 1990s, scientific contributions on psychological testing tended to discount forced-choice formats for the control of response bias and recommend against their use in inventory construction (e.g., Anastasi & Urbina, 1997; see also Paulhus, 1991). In addition, use of forced-choice items in commercial personality inventories is relatively uncommon. Recent studies reconsider forced-choice format and point out the increasing validity for applicant personality assessment (Bartram, 2007; Christiansen, Burns & Montgomery, 2005).

Nevertheless forced-choice formats are seldomly used in a practical context. For example, Goffin and Christiansen (2003) recently reviewed the strategies used to combat motivated distortion in 14 of the personality inventories most commonly used in applied settings. Of these, only one test relied on forced-choice items exclusively (the Occupational Personality Questionnaire 4.2; SHL, 1998) and one for approximately 10% of the items (the PDI Employment Inventory; Paajanen, Hansen, & McLellan, 1993). The remaining inventories all use normative, single- stimulus items. This research takes a closer look at forced-choice formats and addresses the reliability of self-evaluation data focusing on two main research questions:

The first question focuses on the reliability of self-evaluation depending on the rating format, the second question focuses on the reliability depending on the item format.

*Reliability depending on the rating format*

A personality, motivation or interest questionnaire is said to be ipsative when the sum of the scores obtained over the attributes of scales measured for each respondent is a constant (Clemens, 1966). Example item(s) scored as an ipsative quad and normatively

| | ipsative | | normative | | |
|---|---|---|---|---|---|
| | Most true | Least true | not true | true | very true |
| (a) I treat customers pro-actively and friendly | 0 | 0 | 0 | 0 | 0 |
| (b) I cope easily with difficult circumstances | 0 | 0 | 0 | 0 | 0 |
| (c) I take the perspective of others | 0 | 0 | 0 | 0 | 0 |
| (d) I produce many imaginative ideas | 0 | 0 | 0 | 0 | 0 |

Table 1: Comparison of ipsative and normative measurement

In the ipsative case a respondent is asked to indicate which of the four options are most and least true. Clearly the choice of 'I treat customers pro-actively and friendly' as most true (score +1) produces a positive score on one scale (say 'customer orientation'), but also causes lower scores to be obtained on the scales represented by the other three options (score either 0 or, if the least true is chosen, - 1). Because a respondent assigns two 0 responses, a + 1 and a - 1 in each quad, the total score which that respondent receives overall will be the same (i.e. zero), but the scores will be distributed differently across the scales, depending on the choices which that individual has made. It is this which causes mathematical dependence between scales and has given rise to the generalization that ipsative scoring cannot be used for comparing between individuals.

With normative scaling the respondent selects one option from a range of graded responses to each item. Response alternatives may be not true – true - very true, like – indifferent - dislike, or a four-, five- or even seven-point rating scale from say 'strongly agree' to 'strongly disagree'. As each item is endorsed separately the assumption is made that the normative format does not produce mathematically interdependent scores and is therefore superior to the ipsative. Why use ipsative scaling? Ipsative scaling is used for two main reasons: for better control of response sets, such as rating tendencies or social desirability and to make Self- and other ratings more comparable eliminating self enhancing tendencies (Pronin E, Gilovich T & Ross L., 2004) of self ratings. In addition the authors of the present study could show in previous empirical work, that according to the “bottleneck" conception of human information processing (e.g., Broadbent, 1958; Treisman, 1969), raters show great difficulties integrating absolute competency level and profile information when completing Likert type questionnaires used in normative tests. These reasons

favouring a forced-choice format are discussed later in this paper once the results on reliability of forced-choice format have been reported.

One of the main objectives of this paper is more of methodological nature showing that ipsative rating format produces interpretable and reliable data. The authors of the present study suggest that a forced-choice format provides a solid instrument to assess self-rated competency profiles with reliability coefficients that can easily compete with reliability scores of normative instruments. As a benchmark the most common and widely spread personality questionnaires using a Likert-scale item format are considered. We cite the German BIP – Bochum Inventory for job-related Personality description (Hossiep & Paschen, 2003) with a test-retest reliability of r=.83, the Revised NEO Personality Inventory (Costa & McCrea, 1992) with a test-retest reliability of r=0.86, the Californian Personality Inventory (Gough, 1987) with a retest coefficient of 0.84, the Eysenck Personality Inventory (Eysenck & Eysenck, 1975) with a coefficient of r=0.81 and the Sixteen Personality Factor Questionnaire (16 PF) Cattell et. al (1995) with a reliability score of r=0.80. Loewenthal (1996) suggests that for scales with ten or fewer items a reliability of 0.6 is acceptable. The test-retest scores of the here quoted personality instruments all show acceptable values indicating a good level of test-retest reliability with reliability scores between 0.81 and 0.86. Other competency inventories such as the EQ-i has shown adequate test–retest reliability of 0.85 after 1 month and 0.75 after 4 months (Bar-On, 1997).

Based on the research outlined previously, the following hypothesis was developed:

H1: Forced-choice formats can reach equivalent test- retest reliability as normative rating formats.

According to the procedure of traditional test construction of normative personality inventories, we also looked at the reliability scores on construct level. The assessment of scale reliability is based on the correlations between the individual items or measurements that make up the scale, relative to the variances of the items. Often up to six or seven items are used to measure a specific theoretical construct such as management competencies. Namely, the more items there are in a scale designed to measure a particular construct, the more reliable will the measurement (sum scale) be. Since we have only 1 item for each construct, we don’t except high reliability scores at construct level in the two measurement points.

*Reliability depending on the item format*

The second question focuses on the reliability of ipsative measurement depending on the item format. Although research has shown that differences in item format have only minimal effects on the quality of ratings (Landy & Farr, 1980), different formats involve different psychological processes (see Murphy & Constans, 1987) that may influence the relationship between individual differences or contextual variables and ratings. Specifically, item formats vary in the degree to which they structure the rating task, and thus require different levels of cognitive processing (Murphy&Cleveland, 1995).

To the knowledge of the authors, the item format has only been examined on the basis of Likert-typed scales used in performance appraisal systems. Yun, Donahue, Dudley and Mc Farland (2005) have examined a graphic rating scale and behavioural checklist as two different item formats. In their practical implications they recommend to use behavioural checklists in order to simplify the evaluation process by reducing it to a near objective level. These findings correspond to the findings in assessment literature stating that behavioural checklists may assist raters to observe specific behaviours, and provide retrieval cues in recalling behaviours (Donahue et al., 1997; Reilly et al., 1990).

In the present study the authors used two different item formats that were applied in a forced-choice setting. The first item format represents a behavioural anchor of a competency describing the specific meaning of a competency in a practical context. The second item format represents only a specific term encoding the behaviour without triggering a specific situational context. The assumption is that a single term representing a certain category of behaviour facilitates the conception of cross situational traits. The interest in this question and its relevance for the construction of competency assessment using a forced-choice format has arisen with the controversial question whether personality traits truly exist or are merely artefacts in the minds of observers (e.g.. Fiske, 1978, Mischel, 1968, 1973). Hogan's (2005) statement “If you don’t like personality measures, what are the alternatives?” in his somewhat provocative article *In Defense of Personality Measurement: New Wine for Old Whiners* shows that this debate is not at its end. Of the many uses of the trait concept, perhaps the most critical is the assumption that the numerous manifestations of a person’s behaviour can be subsumed by underlying stabilities in character, personality

or even competency. Thus the many ways of expressing extraversion or openness to experience might be traced to the existence of a stable construct called extraversion or openness to experience. Without this assumption, the trait concept loses much of its scientific appeal. Writers such as Mischel have thrown considerable doubt on the assumption of stable traits by appealing to the situation specificity and plasticity of behavioural manifestations of traits. Others (e.g., Bowers, 1973; Endler & Magnusson, 1976,) have emphasized the importance of the interaction between person variables such as traits and situations. Newer studies have confirmed the situational impact on trait measurement in relation to job performance. (Fleeson, 2007, Barrick, Parks & Mount, 2005, Barrick, Mitchell,
& Stewart, 2003; Hough, 2003; and Judge & Kristof-Brown, 2003). Although the issue is far from settled, attacks have placed the validity of the trait concept somewhat in question. One very strong case against the trait concept can be derived from research on the attributional behaviour of the layperson. A widely observed phenomenon, the fundamental attribution error first introduced by Jones (1979), suggests that the layperson overestimates the contribution of personality characteristics as causes of behaviour and similarly ignores the effects of situations. (e.g. Specht, Fichtel & Meyer, 2007; Ensari & Miller, 2006; Ellis, Ilgen & Hollenbeck, 2006). Thus it can be argued that the scientific appeal of traits actually reflects the layperson's tendency to overattribute underlying stability in others' behaviour when actually behaviour might largely be determined by situations. This error might then be compounded by tendencies to make errorfree predictions of future or related behaviour from small samples of behaviour and to draw these inferences on the basis of heuristics such as representativeness rather than careful attention to actual behaviour rates (Kahneman & Tversky, 1973).

Shweder (1975, 1977) and D'Andrade (1965, 1974) advanced one extremely provocative interpretation of these attribution tendencies. According to their "systematic distortion" position, observers impose semantic structure on behaviour, when in fact no such structure exists. That is, rather than recognizing the true empirical relations between behaviour categories, observers use a representativeness or similarity heuristic to describe the relations between categories. Konstabel & Virkus (2006) could show in their quite recent study that the structure of self-rated traits is not reducible to semantic similarities of traits descriptors. Konstabel & Virkus

argue with Peabody and Goldberg's (1989) principle of cognitive economy, which views internal structure as a simplified representation of the observed facts of life.

We resume this argument by drawing the conclusion that behaviour rating is a complex, but not necessarily reflected process including the latent variables of behaviour, behaviour encoding through semantics and memory of behaviour. Hence, we don't see a heuristic as a biased shortcut, but rather as an effect of a usually non-conscious information integration process.

To encode behaviour such as "negotiate", for example, requires mapping rules between events or states (e.g. stating own opinions, listening to others, focusing on the target, claim results etc.) and the behaviour label. Semantics also define relations between behaviour labels. Thus "negotiating" is perhaps somewhat similar to "presenting", but different to "planning and organizing". Indeed, one might say that the behaviour-encoding process implicitly rests on such similarity judgments because the observer must decide how congruent a given behaviour is with all of the many labels that might be matched to it before he or she arrives at the appropriate description. Although we accomplish this judgment with great ease, it is by no means understood how we do so, and at this point we only conclude that this process occurs. However, an implication of this complex process is that different semantic structures produce different encodings. Thus although we assume that this process is consistent with one of Shweder and D'Andrade's (1979) assumptions that semantics determine how we describe reality, there is nothing magical about this process. Indeed, the effect of semantics is purely definitional in the sense that except for error in perceiving behaviour, semantics cannot be wrong at this stage.

The third latent variable in this model of self-rating is our memory. The memory of the encoding process is presumably the impression that someone has of his own behaviour across different situations. We argue that memory-based ratings of competencies are typically scaled across several behaviour categories. The key question at this point is, whether memory based ratings are of better reliability when we use a label for a certain behaviour category or whether we describe the behaviour with a specific behavioural anchor without a behavioural label. Based on the already introduced argument of our hypothesis is that interbehaviour relations are more reliably measured when the encoding process of behaviour is labelled with one

specific term representing the behaviour category instead of a behavioural anchor without a specific competency term.

H2: The reliability of self-evaluations using a forced-choice format is higher when the items consist of a single label for the behavioural category rather than a behavioural anchor of the specific behaviour without a competency label.

## 9.2 Method

*Participants*

The participants of the study were 28 managers (middle and upper management) employed in a large utility company in Switzerland. The sample was heterogeneous in gender, job function and hierarchy level.

*Procedure*

The participants were told the study`s aim: To investigate the effectiveness of a new method of competency assessment. Detailed information regarding the definition of each dimensions used and elaborate instructions on how to complete the online questionnaire based on a forced-choice mechanism for all the 75 items were provided. The 75 items consisted of 15 competencies and 60 competency facets corresponding to the 15 competencies. Participants could take as long as needed to respond to each item.

In both measurement points (t1 and t2) participants were asked to do a self rating based on the 15 competencies. In addition all the 60 competency facets were presented. In the first measurement point the 60 competency facets were presented as behavioural anchors including the label of the competency (in this study labelled as BA). In the second measurement point the 60 competency facets were presented as mere competency labels without any description or definition (in this study labelled as CL).

The rating procedure included two steps. In a first step raters were ask to categorize the items (15 competencies at a time) into three categories. Each category allowed only a certain amount of competencies.

- very high developed competency (max. 7 competencies)
- high developed competency (max. 4 competencies)
- less high developed competency (max. 4 competencies)

In a second step the participants were asked to rank the competencies within each category according to their self-perception of their own strengths and weaknesses based on the given competencies. This methodology is similar to the ideal point models used in marketing research. (e.g. Green & Carmone, 1969, Schaupp & Belanger, 2006, Gustaffson, Hermann & Huber, 2007; Baier & Gaul, 2007) whereas the competence order refers to rank order of preference data.

There were two measurement points with a time interval of two to maximum three weeks. In the first measurement point raters were asked to do the procedure described above with 15 competencies defined with a competency label and the corresponding behavioural anchor. Then 60 competency facets followed described only with the behavioural anchor. Due to reasons of a cognitive bottleneck in human information processing (e.g., Broadbent, 1958; Treisman, 1969) only 15 items (4 x 15) were presented at the time. In the second measurement point the same 15 competencies were presented in a first step and in a second step the same 60 competency facets (4 x 15 items) were presented, this time only with a competency label without the corresponding behavioural anchor.

Following this procedure the researchers of this study obtained for each individual a competency profile, including all the different items, with a numerical value between 1 and 15 according to the rank order.

*Design*

In order to test the two hypothesis stated above, a 2 x 2 design with two measurement points with a time interval of 2 to 4 weeks and two different item formats was used: *competency label (CL)* and the *behavioural anchor(BA)* of a competency without a specific label.

| | **Measurement point** | |
|---|---|---|
| | **t1** | **t2** |
| **Item format** | 15 Competency dimensions with definition, measurement point (t1) | 15 Competency dimensions with definition, measurement point (t2) |
| | Behavioural anchors for 60 competency facets (BA) | Competency label for 60 competency facets (CL) |

Table 2: Design for reliability testing depending on item format and measurement point

*Measures*

In order to test the first hypothesis of reliability, self-ratings of the first and second measurement point were compared. Since in ipsative measurement the sum of the scores obtained over the attributes or scales measured is a constant for each individual, ipsative measurement has been criticized for the use of interindividual comparison. In this study reliability has not been measured on a scale by scale basis, but on a holistic level using Nonmetric Multidimensional Scaling (NMDS) comparing individuals based on their self rated competency profiles. Methods of multidimensional scaling or ordination seek a parsimonious representation of individuals in a space of low dimensionality. Parsimony in this context implies that the distances between individuals in ordination space optimally represent their dissimilarities in variable space, in some defined sense. Techniques differ in their definition of optimality. But a minimal requirement of most methods is a rank order agreement between distances and dissimilarities (Shephard and Carroll 1966, Orloci 1978).

The NMDS maps used in this study have been computed using ROBUSCAL technique which has been proven to be a robust scaling technique minimizing the effect of outliers in a specific data set. Computing NMDS based on Pearson Correlations we obtain two dimensional maps representing self-rated competency profiles at two different measurement points. Analysing the change in position of the individuals according to the two different measurement points lead to concise conclusions in terms of the reliability of a test instrument.

In other words this means that the applied forced-choice instrument is only to be considered as a reliable instrument, if the position of the competency profiles in the NMDS map of measurement point t1 and t2 reveals to be relatively stable.

In order to quantify the reliability of the forced-choice instrument we introduce three different testing models, which increase in terms of their testing severity:

1. Distributional-level effect model: The mean Pearson correlation between t1 and t2 over all individuals (correlation for each individual between measurement point 1 and 2) is significantly higher than the **mean** inter-correlation among the raters in t1 and t2.

2. Individual-level effect model: For each single individual, the Pearson correlation of t1 and t2 is higher than the **maximum** inter-correlation with any other profile of a different individual within t1 and t2.
3. Structural-level effect model: For each individual, the pair-wise distances of self-ratings in t1 and t2 (as a result of the extracted distances of the NMDS using Pearson correlation between t1 and t2) are smaller than a critical value defined as the mean distance minus one standard deviation of the NMDS map of t1 and t2.

The same models of testing reliability were also applied to test the second hypothesis regarding the item format of behavioural anchors and competency labels.

If we are interested in the reliability of a test instrument depending on the rating format and the item format, we can also apply the method of Procrustes transformation of NDMS maps. The Procrustes transformation compares two Euclidian maps by extending, shifting, rotating and mirroring the configurations to approach a maximal congruence and then determines the remaining deviation as the average loss between the Euclidian maps. Average loss, defined as the mean distance of two corresponding items in the two maps, have a value between 0 and approximately 2. If the average loss is 0, the two configurations of NMDS are absolutely identical. If the average loss is >1 the two structures of the NMDS maps have not much in common and have to be considered as stochastically independent (see Läge, 2001). The average loss in the context of this study gives an indication on how much the structure of two different NMDS has changed. In this study we used Procrustes transformation to measure the change according to the 2 x 2 design, in order to get an understanding on how the reliability changes depending on the time influence and the influence of the different items used in the forced-choice format.

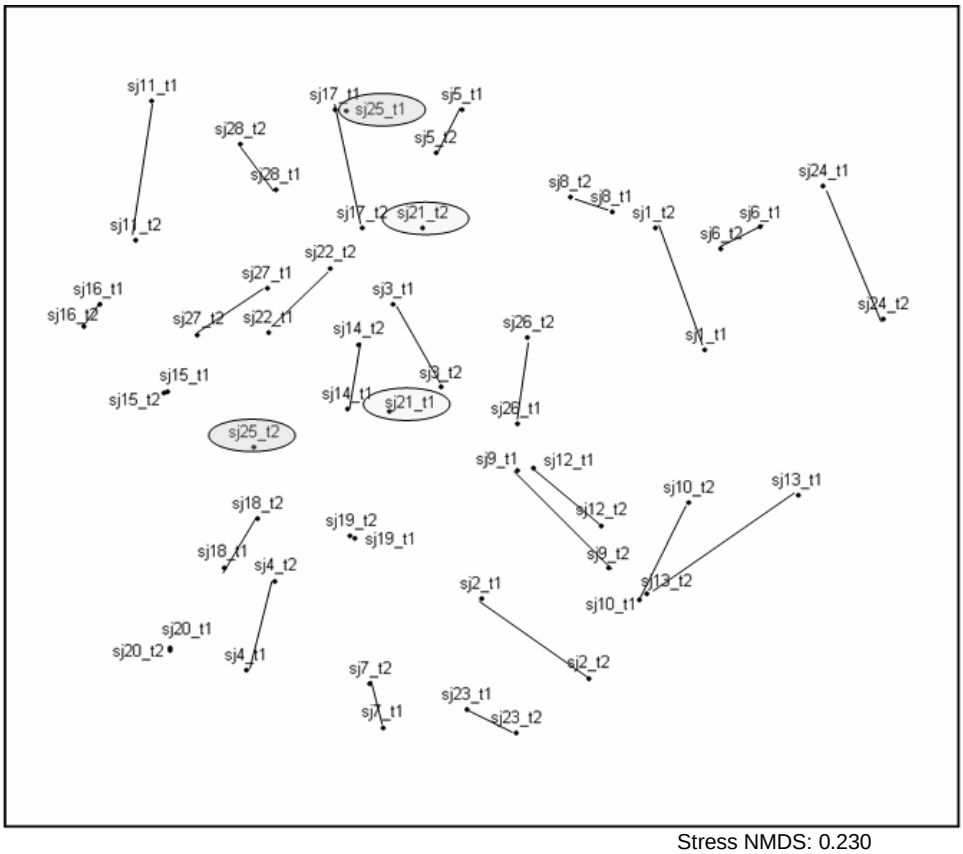
## 9.3 Results

When looking at the data with classical Test-Retest correlation coefficients, high reliability scores can be reported. The mean correlation between the self ratings of the first measurement point (t1) and the second measurement point (t2) across all dimensions is 0.8 with a standard deviation of 0.10 which easily equals to scores found in literature of normative measurement stated previously.

In order to get a deeper understanding of what this correlation coefficient means, we also calculated a NMDS map with the data of the two different measurement points.

Figure 1 outlines the two-dimensional Euclidian map of the NMDS based on Pearson correlations. This map depicts the relational structure of the self concepts of individuals based on self assessed competency profiles. Thus, the dots represent individual competency profiles of 28 forced-choice self ratings at two measurement points (t1 and t2). The connection line between two dots illustrates the "movement" of the position of each profile according to the change due to the two different measurement points. This change in structure is a more visual measure for the reliability of a test instrument than commonly used correlation coefficients.

Before looking at the map with regard to content, it is necessary to explain an internal quality measure of the ROBUSCAL algorithm: the standardized stress value (Läge, 2001) is a measure of how well the algorithm was able to translate the similarity judgments into an n-dimensional map. Thus, it is also an indicator of the level of consistency of the ratings across the different scales given by the participants throughout the online self assessment and consequently provides a marker of the interpretability of the cognitive map. According to the literature (Borg & Grönen, 1997; Gigerenzer 1981) the stress value of this mean map of 0.23 is with an object number of 56 (2 x 28) more than acceptable, and thus interpretable.




Stress NMDS: 0.230

Figure 1: NMDS map with all the self rated competency profiles of two measurement points (t1 and t2)

The NMDS map shows a relatively stable position of the individuals (self ratings) across two different measurement points. There are only two individuals (Sj25 and Sj21) who's self-rating differs relatively strong from measurement point 1 to measurement point 2. The results of the NMDS solution reveal high reliability in self-ratings through a forced-choice format.

In the previous chapter we have introduced different testing models in order to quantify the reliability of the forced-choice data in a more differentiated way than just by reporting a correlation coefficient across all individuals. Table 4 illustrates the figures for the different reliability testing models:

| | **Distributional level effect model** Corr. ba and cl > mean inter-corr. of ba and cl | **Individual level effect model** Corr. ba and cl > max. intercorr. within ba and cl | **Structural level effect model** Distances NMDS and t2 < critical value 0.74 |
|---|---|---|---|
| sj1 | **0.43 > 0.14** | **0.43 > 0.41** | **0.67 < 0.73** |
| sj2 | **0.23 > 0.14** | *0.23 < 0.34* | *0.92 >0.73* |
| sj3 | **0.53 > 0.14** | **0.53 > 0.34** | ***1.24 > 0.73*** |
| sj4 | **0.53 > 0.14** | **0.53 > 0.39** | **0.51 < 0.73** |
| sj5 | **0.54 > 0.14** | **0.54 > 0.49** | **0.12 < 0.73** |
| sj6 | **0.62 > 0.14** | **0.62 > 0.38** | **0.46 < 0.73** |
| sj7 | **0.52 > 0.14** | **0.52 > 0.39** | **0.13 < 0.73** |
| sj8 | **0.72 > 0.14** | **0.72 > 0.38** | **0.26 < 0.73** |
| sj9 | **0.31 > 0.14** | **0.31 > 0.25** | **0.53 < 0.73** |
| sj10 | **0.63 > 0.14** | **0.63 > 0.39** | **0.35 < 0.73** |
| sj11 | **0.48 > 0.14** | *0.48 < 0.51* | **0.53 < 0.73** |
| sj12 | **0.46 > 0.14** | **0.46 > 0.35** | **0.31 < 0.73** |
| sj13 | **0.38 > 0.14** | **0.38 > 0.31** | **0.61 < 0.73** |
| sj14 | **0.52 > 0.14** | **0.52 > 0.37** | *1.53 > 0.73* |
| sj15 | **0.69 > 0.14** | **0.69 > 0.51** | **0.14 < 0.73** |
| sj16 | **0.40 > 0.14** | **0.40 > 0.37** | *1.42 > 0.73* |
| sj17 | **0.57 > 0.14** | **0.57 > 0.41** | **0.26 < 0.73** |
| sj18 | **0.56 > 0.14** | **0.56 > 0.36** | **0.40 < 0.73** |
| sj19 | **0.52 > 0.14** | **0.52 > 0.45** | **0.58 < 0.73** |
| sj20 | **0.31 > 0.14** | *0.31 < 0.35* | *1.02 > 0.73* |
| sj21 | **0.48 > 0.14** | **0.48 > 0.31** | **0.31 < 0.73** |
| sj22 | **0.23 > 0.14** | *0.23 < 0.45* | **0.21 < 0.73** |
| sj23 | **0.50 > 0.14** | **0.50 > 0.39** | *0.78 > 0.73* |
| sj24 | **0.42 > 0.14** | **0.42 > 0.41** | **0.25 < 0.73** |
| sj25 | **0.53 > 0.14** | **0.53 > 0.34** | **0.40 < 0.73** |
| sj26 | **0.33 > 0.14** | **0.33 > 0.32** | **0.64 < 0.73** |
| sj27 | **0.59 > 0.14** | **0.59 > 0.44** | **0.30 < 0.73** |
| sj28 | **0.50 > 0.14** | **0.50 > 0.49** | **0.53 < 0.73** |

Table 3: Three different models of testing reliability of a forced-choice rating instrument

The first column represents the distributional level effect model for reliability. The correlations coefficients were calculated by using Pearson correlations for each

subject's competency profile of the measurement point t1 and t2. The distribution of the correlations of t1 and t2 has then been compared with the distribution of the mean inter-correlations between the subjects of t1 and t2. The mean of the inter-correlations of t1 and t2 is 0.10 with a standard deviation of 0.27. A quick glance at the correlations coefficients is sufficient to realize that Test-Retest reliability should be on an acceptable level. When computing a t-Test of the correlations coefficients of t1 and t2 with all the inter-correlations within t1 and t2, the mean difference is highly significant ($p<0.01\%$) with an effect size of $d=3.43$.

In the second column the correlations between t1 and t2 are compared with the maximum inter- correlations among the participant's competency profiles of t1 and t2. We assume that there is a systematic connection between the competency profiles of measurement point 1 and measurement point 2, if the correlation coefficients are generally higher than the highest inter-correlations of the competency profiles among the raters within the two different measurement points. When we look at the scores, we can report 26 out of 28 correlations-pairs between t1 and t2 which are higher than the maximum inter-correlations within measurement point t1 and t2. This finding represents a solid indicator for a systematic connection between the two measurement points.

The third column depicts the *distances NMDS t1 and t2* and illustrates the toughest model for testing reliability. Taking into consideration the relationship between the different profiles, we measure the stability of the forced-choice instrument on a structural level. The distances were calculated from the NMDS map by extracting all the distances from the map illustrated in figure 1, which is based on Pearson correlations between all the self rated competency profiles through a forced-choice mechanism described earlier in this paper. High reliability is given, when the distances between t1 (competency profile of measurement point t1) and t2 (competency profile of measurement point t2) are small. In order to define "small distance" we introduce a critical criteria defined as the mean distance minus one standard deviation of all the distances in the NMDS. The criterion for small distance is therefore 0.74 (mean 1.39 – standard deviation 0.65). When applying this criteria, only two (VP 21 and VP 25) out of twenty eight participants do not correspond this criteria by exceeding the "small distance" of 0.74 in the NMDS map.

After this differentiated analyses we can definitely confirm our first hypothesis. A self-reported competency profile based on a forced-choice format produces equivalent reliability scores to normative personality inventories.

In order to illustrate the reliability at construct level, we calculated Pearson correlations for each construct over two measurement points (see table 4).

| | | | | |
|---|---|---|---|---|
| **Entrepreneurship** | **Problem solving** | **Planning & Organ.** | **Open to experience** | **Customer Orient.** |
| 0.88 | 0.83 | 0.85 | 0.62 | 0.88 |
| **Conflict resolution** | **Resilience** | **Selfmanag.** | **Achievement** | **Quality awareness** |
| 0.73 | 0.67 | 0.75 | 0.68 | 0.83 |
| **Communic.** | **Teamwork** | **Inspiring others** | **Leadership** | **Develop others** |
| 0.79 | 0.75 | 0.84 | 0.59 | 0.82 |

Table 4: Reliability coefficients on construct level

Most of the competency concepts show high reliability over all 28 participants. Only Leadership and open to experience have correlation coefficients below r=0.65. The other competency constructs show correlation coefficients that are on a comparable level to traditional personality and competency tests (for an overview, see also Boyle, Matthews and Saklofske, 2008).

The significance of these results and its practical implications will be discussed in the last chapter of this paper.

After having reported the results according to our first hypothesis we now present the results that examine our second hypothesis which focuses on the item format of forced-choice measurement. As mentioned earlier we are interested weather the reliability is dependent on the item format. We applied the same methodology as in the reliability testing of the two different measurement points.

If we look at the reliability score of the behavioural anchors (items used in measurement point 1) and the competency labels (items used in measurement point 2), we observe a decrease in reliability of roughly 40%. The mean Pearson correlation over all individuals is 0.49, with a standard deviation of 0.12.

In order to obtain a more differentiated view on this decrease in reliability, we illustrate in Figure 2 the NMDS map of the 28 competency profiles that haven been measured based on 60 competency facets described with behavioural anchors and 28

competency profiles that have been measured based on the same 60 competency facets but described only through a specific label for the relevant competency.

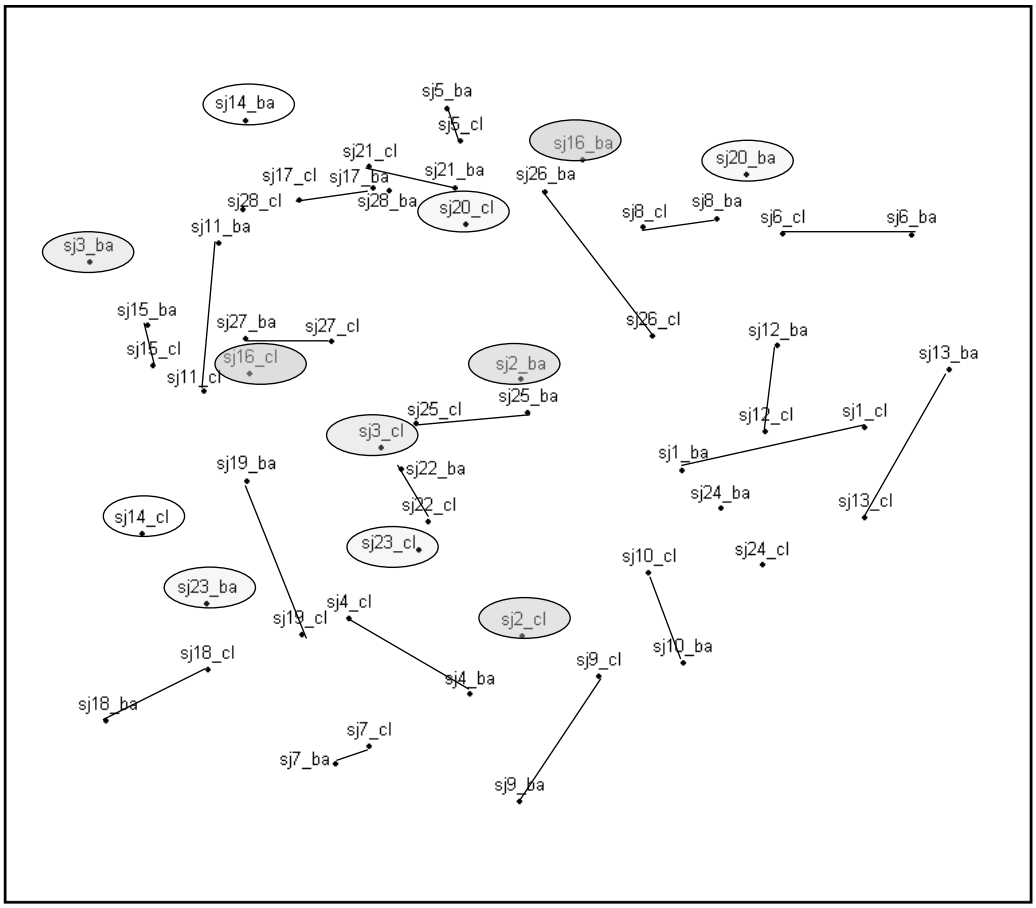



Figure 2: NMDS map with all the self rated competency profiles based on behavioural anchors (ba) and competency labels (cl)

Before looking at the map with regard to content, the stress value of 0.27 is pointed out which according to Gigerenzer 1981 is a decent value considering the amount of objects and the number of dimensions in the NMDS ($p<0.01$ in his Monte Carlo study for n=56 and dimensions = 2). Thus, the map is interpretable.

The position of the competency profiles indicates a vaster change in position depending on the item format. However, the relational structure of the map shows that the position of most competency profiles does not change dramatically which indicates that the influence of the item format has not a decisive impact on the reliability illustrated in the NMDS map.

For more detailed analysis of these results the same three models of testing for reliability as for testing K1 and K2 were applied for quantitative testing.

| | **Distributional level effect model** Corr. ba and cl > mean inter-corr. of ba and cl | **Individual level effect model** Corr. ba and cl > max. intercorr. within ba and cl | **Structural level effect model** Distances NMDS and t2 < critical value 0.74 |
|---|---|---|---|
| sj1 | **0.43 > 0.14** | **0.43 > 0.41** | **0.67 < 0.73** |
| sj2 | **0.23 > 0.14** | *0.23 < 0.34* | *0.92 >0.73* |
| sj3 | **0.53 > 0.14** | **0.53 > 0.34** | ***1.24 > 0.73*** |
| sj4 | **0.53 > 0.14** | **0.53 > 0.39** | **0.51 < 0.73** |
| sj5 | **0.54 > 0.14** | **0.54 > 0.49** | **0.12 < 0.73** |
| sj6 | **0.62 > 0.14** | **0.62 > 0.38** | **0.46 < 0.73** |
| sj7 | **0.52 > 0.14** | **0.52 > 0.39** | **0.13 < 0.73** |
| sj8 | **0.72 > 0.14** | **0.72 > 0.38** | **0.26 < 0.73** |
| sj9 | **0.31 > 0.14** | **0.31 > 0.25** | **0.53 < 0.73** |
| sj10 | **0.63 > 0.14** | **0.63 > 0.39** | **0.35 < 0.73** |
| sj11 | **0.48 > 0.14** | *0.48 < 0.51* | **0.53 < 0.73** |
| sj12 | **0.46 > 0.14** | **0.46 > 0.35** | **0.31 < 0.73** |
| sj13 | **0.38 > 0.14** | **0.38 > 0.31** | **0.61 < 0.73** |
| sj14 | **0.52 > 0.14** | **0.52 > 0.37** | *1.53 > 0.73* |
| sj15 | **0.69 > 0.14** | **0.69 > 0.51** | **0.14 < 0.73** |
| sj16 | **0.40 > 0.14** | **0.40 > 0.37** | *1.42 > 0.73* |
| sj17 | **0.57 > 0.14** | **0.57 > 0.41** | **0.26 < 0.73** |
| sj18 | **0.56 > 0.14** | **0.56 > 0.36** | **0.40 < 0.73** |
| sj19 | **0.52 > 0.14** | **0.52 > 0.45** | **0.58 < 0.73** |
| sj20 | **0.31 > 0.14** | *0.31 < 0.35* | *1.02 > 0.73* |
| sj21 | **0.48 > 0.14** | **0.48 > 0.31** | **0.31 < 0.73** |
| sj22 | **0.23 > 0.14** | *0.23 < 0.45* | **0.21 < 0.73** |
| sj23 | **0.50 > 0.14** | **0.50 > 0.39** | *0.78 > 0.73* |
| sj24 | **0.42 > 0.14** | **0.42 > 0.41** | **0.25 < 0.73** |
| sj25 | **0.53 > 0.14** | **0.53 > 0.34** | **0.40 < 0.73** |
| sj26 | **0.33 > 0.14** | **0.33 > 0.32** | **0.64 < 0.73** |
| sj27 | **0.59 > 0.14** | **0.59 > 0.44** | **0.30 < 0.73** |
| sj28 | **0.50 > 0.14** | **0.50 > 0.49** | **0.53 < 0.73** |

Table 5: Three different testing models for reliability between behavioural anchor and competency label

Looking at the first column, the correlations coefficients vary between 0.23 and 0.73 with a mean of 0.48 and a standard deviation of 0.12. These correlations coefficients for each individual are significantly higher than the mean inter-correlation coefficient ($\bar{r}$ = 0.14) of the two distributions of behavioural anchors, measurement point 1, and the competency labels, measurement point 2, (p<0.01%, d=2.50).

When applying the tougher criterion of the highest inter-correlations within the two different item formats distributions, 4 out of 28 participants have any higher correlation with the profile of a different subject than between the behavioural anchors and the competency labels. All other 24 participants cannot be mistaken by a different subject when there is a modification in the item format. The majority of the

participants show a higher correlation between the two measurement conditions with a different item format.

When comparing the distances extracted from the NMDS map and challenging the values with the defined criterion of 0.73 (mean distance of 1.37 – 1 standard deviation of 0.64 = 0.73), only 6 out of 28 distances do not comply this criterion of a small distance. The majority of the pair wised distances of the behavioural anchor item format and the competency label item format in the NMDS map can be considered as small. In regard to content this means that participants can produce a relatively stable self concept regardless of the item format.

In order to explain the decrease in reliability when considering the reliability results of t1 and and t2 (1st hypothesis) compared to the reliability of the different item format (2nd hypothesis), we present the results of Procrustes Transformation. This method, as described earlier in the methodological part of this paper, is a similarity measure to compare the structure of two different NMDS maps. By comparing the different NMDS maps in the different measurement conditions controlling for time effect and item effect, we try to understand which mechanisms lead to a decrease of reliability in our forced-choice instrument.

We introduce the Procrustes transformation showing the two NMDS maps from measurement point t1 with the 60 behavioural anchors as items and measurement point t2 with the 60 competency labels as items and then in third map we show the Procrustes solution revealing the change in structure of the two maps.

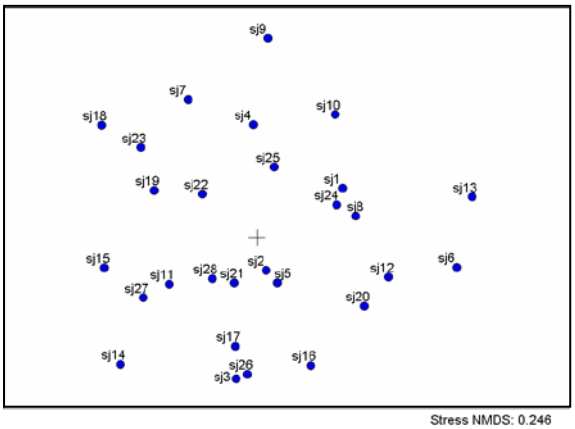



Figure 3: NMDS map behavioural anchor

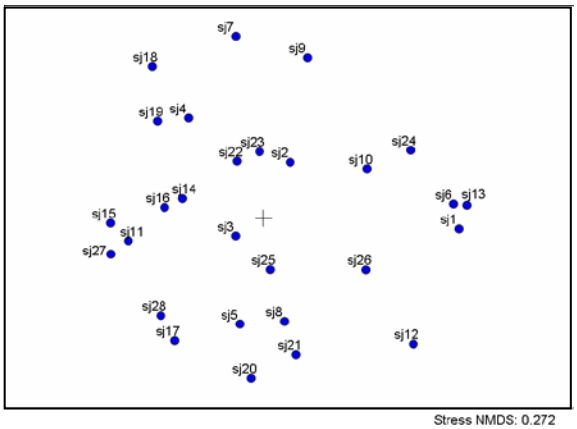



Figure 4: NMDS map competency labels

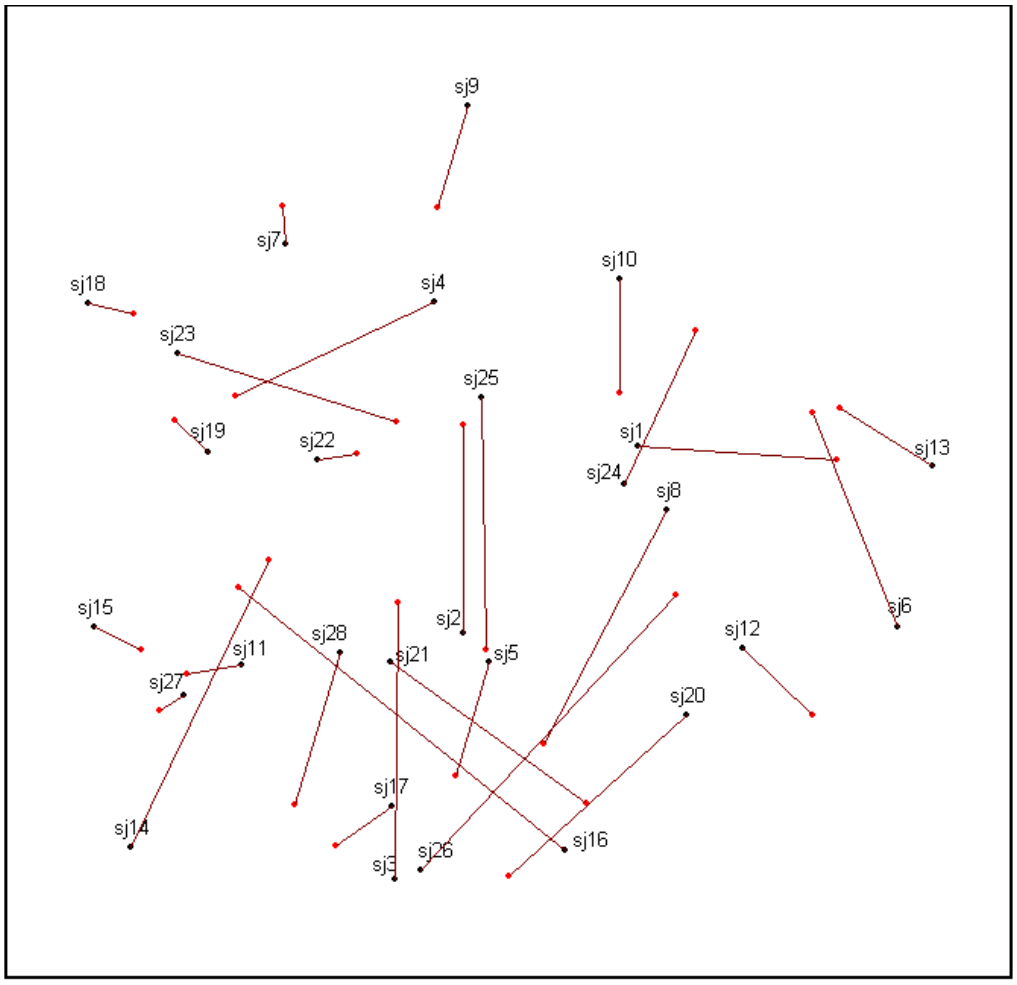



Figure 5: Procrustes map of behavioural anchors and competency labels

The Procrustes transformation shows an Average Loss of only 0.43. (The Average Loss equals the averaged Object Losses. A value of 1.00 would represent the distance between any corresponding objects being equal to the mean distances within the two NMDS maps). Hence, the resulting NMDS structures are similar, and not much more dissimilar than the two NMDS structures of competencies at t1 and t2 (Average loss = 0.38). Only 5 individuals out of 28 show higher Object losses than 0.65. In conclusion to our second hypothesis we can say, that the item format has not much effect on the structures of profiles, despite the correlations have been proven to be much more moderate than the high correlations of the test-retest reliability of t1 and t2.

According to the illustrated Procrustes procedure we computed the Average Losses for the different measurement conditions.

Table 6 shows the average losses of the NDMS maps related to the different measurement conditions illustrating the influence of time (measurement point t1 vs.

t2), the influence of item format (behavioural anchor vs. competency label), as well as the influence of the correspondence between behavioural anchor and competency label.

| Effect | Description of procrusted NMDS maps | Avg Loss |
|---|---|---|
| time | 15 Comp. t1 and 15 Comp. t2 | 0.38 |
| item format as behavioural anchors | 15 Comp. t1 and 60 behavioural anchors t1 | 0.58 |
| item format as competency labels | 15 Comp. t2 and 60 competency labels t2 | 0.43 |
| correspondance two item formats | 60 behavioural anchors t1 and 60 competency labels t2 | 0.43 |

Table 6: Average Loss to illustrate reliability over time and item format

The results in table 6 reveal that the main structural change of the two different NMDS maps is based on the item format when using 60 behavioural anchors as stimulus material compared with the map of 15 competency labels in t1. The respondents` profiles produce more different Euclidian maps when the items are presented as behavioural anchors (AvgLoss = 0.58). That this effect is not simply due to the different amount of items, is indicated through the relatively low average loss of 0.43 in the condition t1 with 15 competency labels and in comparison with the 60 competency labels also measured in t1. We can also exclude that the effect of the item format is due to lack of correspondence of the behavioural anchors and the competency labels. When comparing the structural change of the Euclidian maps of behavioural anchors and competency labels with the two maps of t1 and t2, there is only a marginal change in average loss (AvgLoss t1 and t2 of 0.43- AvgLoss of BA and CL of 0.38 = 0.05). Later in the discussion chapter we postulate a simple pathmodel explaining the increase of Average Loss when comparing the different measurement conditions.

## 9.4 Discussion

Although the effects of applicant distortion continue to be debated in academic circles, substantial skepticism exists in industry regarding the use of self-report measures to facilitate hiring decisions. Concern that motivated applicants can easily distort personality measures remains the most widespread criticism organizational decision makers have of personality testing (Cook, 1993; Hogan & Hogan, 1992; Hogan et al., 1996). Disregarding such admonitions, many personality inventories popular in industry have been developed without concern over socially desirable

responding due to the belief that self-enhancement represents valid trait variance and, therefore, is not a problem (see e.g., Costa & McCrae, 1992; Hogan & Hogan, 1992). By better addressing these concerns, applied psychologists may engender more confidence in the professionals our practice serves. Indeed, a recent survey of practitioners who work in the area of selection and assessment found that approximately 70% expressed preference for using a personality inventory that includes a method to deal with applicant distortion (Goffin & Christiansen, 2003).

Much has been written about the difficulties and limitations of using ipsative measurement and forced-choice response styles in the assessment of personality using multi-scale questionnaires (Baron, 1996). Baron comes to the conclusion that with a sufficient amount of scales (around 30 scales), ipsative measurement does provide some interpretable psychometric parameters. The present study has shown that also with fewer scales (15 used in this study), a forced-choice format is, from the perspective of reliability, a veritable alternative to Likert type formats found in most normative instruments. With our two- step procedure of putting 15 competencies into a ranking order, we have established a very economic self-assessment tool that is based on a forced-choice format. Instead of long questionnaires with often more than two hundred items used in traditional personality inventories, we introduced a forced-choice procedure that can be completed in a fractional amount of time. Interested in the question whether we reach the same reliability score as normative instruments, we could demonstrate with test-retest correlation coefficients of r=0.8 that we can easily compete with Likert-type scales of normative instruments varying from 0.8 to 0.86. By introducing three models of reliability testing we went beyond the commonly used correlations coefficients. With the *individual level effect model* we could demonstrate the identity of competency profiles on an individual level at two different measurement points. 26 out of 28 individuals cannot be mistaken by a different subject when comparing the competency profile of measurement point 1 and measurement point 2.

Through scaling of the forced-choice data applying NMDS the reliability of the instrument has not only been tested on an individual level but also on a structural level, displaying the relational structure of several forced-choice profiles and their stability over time visualised in two dimensional Euclidian maps. This most severe reliability testing model presented in this study has shown that 26 out of 28 distances,

i.e. difference in competency profiles measured through forced-choice, between measurement point 1 and 2 are smaller than the mean distance of the NMDS map minus one standard deviation. At last, this very severe reliability testing criterion indicates high reliability of the forced-choice method applied in this study.

That these results of reliability are quite remarkable is emphasized through the fact that in this study only a small amount of items has been used without previous item analysis through principal component analysis. Part of the high reliabilities found in classical normative instruments are due to several items that measure the same construct and therefore inflate the reliability measure through correlation measures.

We believe that our high reliability scores are partly due to the intuitive two step procedure of creating a competency profile through forced-choice. This process itself facilitates the process of rank ordering, because it does not require keeping 15 items in the working memory and handling them simultaneously when defining the rank of each competency. In addition, the instrument allows respondents to focus only on profile information and not on the simultaneous self-assessment of absolute score and profile information. We believe that with Likert-type scales raters are cognitively over challenged by integrating both, profile information and absolute competency level.

After having discussed the test-retest reliability of the 15 competencies in the two different measurement points, we want to have a closer look at the item format. Using the same methodologies of reliability testing, we could show the correspondence of behavioural anchors and competency labels. Comparing the Euclidian map of the 60 behavioural anchors with the map calculated on the basis of competency labels as stimulus items, we have found fairly high correspondence of the structures of the profiles, despite the correlations have been proven to be much more moderate.

However, the comparison of the two NMDS maps within one measurement point, comparing the configuration of 15 competencies with the configuration of the 60 behavioural anchors, lead to quite different results on a structural level. The authors postulate two different explanations for this effect of reduced reliability. One is the situational impact of behavioural anchors. Through the definition of a specific behaviour respondents are put into a specific situation and respond according to how they would have behaved in that particular situation triggered by the behavioural anchor of the item. The main principle of cross-situational consistency of personality traits is therefore violated and lead to inconsistent self evaluation concepts. Thus even if traits seem to be an inevitable outcome of individual differences, the possibility

remains that such consistencies are eliminated when individuals observe themselves across different situations.

The second explanation goes more into the direction of Shweder and D´Andrade (1979) who assumed that semantics determine how we describe reality. With the item format, containing only behavioural anchors, the part of encoding into a specific label for a certain behaviour is missing which leads to different ratings due to idiosyncratic encoding of the behaviour. In the setting where the respondents are not asked to first encode the behaviour into semantics, the raters are using the items in a semantically consistent manner, their scores display internal consistency. According to this model of behaviour encoding, semantics define the relations between events (e.g., behaviour) and behaviour categories. Because any behaviour has a complete set of semantic relations with every category, semantics partly define the structure of items in personality tests. If we reconsider the figures of the Procrustes transformation we notice an increase of Average Losses in accordance to the transformation of the stimulus. When we compare the Average Loss from the configuration of the 15 competency labels with a short behavioural description and the configuration of the 60 competency facets described with only competency labels, and we compare these two different stimulus formats within one measurement point, we have found an Average Loss of 0.43. When we compare the correspondence of the 60 competency using behavioural anchors with the 60 competency facets as mere competency labels, measured at two different measurement points, we have found again an Average Loss of 0.43 which indicates a loss of information. The biggest step in Average Loss we find when comparing the 15 competencies with the 60 competency facets described only through behavioural anchors within one measurement point. Here the average Loss of the two configurations comes up to 0.58. We postulate a simple path-model describing the loss of information through the change of the item format:

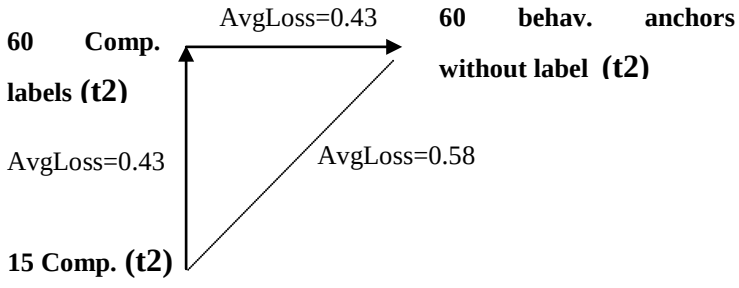



Figure 6: Pathmodel illustrating the Average loss depending on the different item formats

If we interpret this pathmodel we come to the conclusion that behavioural anchors without a specific competency label lead to substantially different self-rated competency profiles.

The practical implication of this interpretation would be that in forced-choice personality tests or competency assessments, researchers and practitioners should be using items that are labelled with a specific competency dimension and a brief definition of what is meant. This format facilitates storage in the working memory and is immune to situational context information influencing the self-rating of competency profiles.

*Conclusions*

To sum it up in a nutshell the authors conclude that individuals can be validly compared on a scale by scale basis on ipsative scales as normative, that ipsative data can be visualised using NMDS and practically tested for reliability showing that reliabilities of ipsative measurement are not overestimated. Based on the results of Baron (1996) who states that “it is unlikely that conclusions with ipsative data, based at least on a relatively large number of scales, are any less valid than those based on the normative”. In extension to the finding of Baron, we could prove that the reliability is not necessarily tied to a large number of scales, but is also given with a small amount of scales, such as 15 competency constructs used in our study. In addition, with ipsative measurement we avoid all difficulties of normative instruments such as acquiescence, social desirability and central tendency. It is true that interpretation of ipsative questionnaires needs to recognize the interdependence which exists between scales, but response sets cause problems of inter-individual comparison with normative scales also. Ipsative scores do have their limitations in terms of the comparison of inter-individual differences on absolute score but so do the normative; it is a matter of trading one type of bias for another. However, researchers and practitioners have to ask themselves the question which purpose they follow when choosing to use a self-assessment instrument. If the purpose is to gain valid profile information on competencies, as required in personnel development, we suggest to use a forced-choice procedure described in this study. For overall competency assessment on absolute score, other predictors such as salary in combination with age might be a better predictor for selection purposes than self-assessed competency profiles.

## 9.5 References

Allport, G.W. & Odbert, H.S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, 47(211).

Anastasi, A., & Urbina, S. (1997). Psychological testing. New York: MacMillan.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, 69, 49–56.

Bar-On, R. (1997). The Emotional Quotient Inventory (EQ-i): a test of emotional intelligence. Toronto: Multi-Health Systems.

Barrick, M. B., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology*, 42, 261–272.

Barrick, M. R. Mitchell, T. R. & Stewart, G. L. (2003). Situational and motivational influences on trait-behaviour relationship. M. R. Barrick & A.M. Ryan (Eds.), *Personality and work: Reconsidering the role of personality in organizations*, 60-82, San Francisco: Jossey-Bass.

Barrick, M. R., Parks, L., & Mount, M. K. (2005). Self-Monitoring as a Moderator of the Relationships between Personality Traits and Performance. *Personnel Psychology*, 58, 745-768.

Bartram, D. (1996). The relationship between ipsatized and normative measures of personality. *Journal of Occupational & Organizational Psychology*, 69, 25–39.

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment*, Vol. 15, Issue 3, 263-272.

Borg, I., & Grönen, P. (1997). Modern Multidimensional Scaling. New York: Springer.

Bowers, K. S. (1973). Situationism in psychology: An analysis and critique. *Psychological Review*: 80. 307-336.

Berkshire, J. R. (1958). Comparison of five forced-choice reference check. *Educational and Psychological Measurement*, 18, 553–561.

Broadbent, D. E. (1958). Perception and communication. New York: Pergamon.

Broverman, C. (1962). Normative and ipsative measurement in psychology. *Psychological Review*, 69(4), 295-305.

Cattell, R.B. (1965). The scientific analysis of personality. Baltimore: Penguin Books.

Cattell, R.B., Cattell, A.K.,& Cattell H.E. (1993). Sixteen Personality Factor Questionnaire, Fifth Edition. Champaign, IL: Institute for Personality and Ability Testing.

Christiansen, N., Burns, G. & Montgomery, G. E, (2005). Reconsidering Forced-Choice item formats for applicant personality assessment. *Human Performance*, 18, 267-307.

Clemens, W. V. (1966). An analytical and empirical examination of some properties of ipsative measures.*Psychometric Monographs*, 14.

Cook, M. (1993). Personnel selection and productivity (Rev. ed.). New York: Wiley.

Cornwell, J. M., & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville and Willson (1992). *Journal of Occupational and Organizational Psychology*, 67, 89–100.

Costa, P. T., Jr., & McCrae, R. R. (1992). The Revised NEO-PI/NEO-FFI manual supplement. Odessa, FL: Psychological Assessment Resources.

D'Andrade, R. G. (1965) Trait psychology and componential analysis. *American Anthropologist*, 67. 215-228.

D'Andrade, R G. (1974). Memory and the assessment of behaviour. In T. Blalock (Ed.), *Measurement in the social sciences* (pp. 159-186). Chicago. Aldine-Atherton.

Dilchert, S., Ones, D. S., Viswesvaran, C., & Deller, J. (2006). Response distortion in personality measurement: Born to deceive, yet capable of providing valid self-assessments? *Psychology Science*, 48, 209-225.

Donahue, L. M., Truxillo, D. M., Cornwell, J. M. & Gerrity, M. J. (1997). Assessment Center Construct Validity and Behavioural Checklist: Some Additional Findings. *Journal of Social Behaviour and Personality*, 12, 85-108.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). As assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance*, 16, 81–106.

Edwards, A. L. (1959). Edwards Personal Preference Schedule manual. New York: Psychological Corporation.

Endler, N. S., & Magnusson, D. (1976) Interactional psychology and personality Washington, DC: Hemisphere.

Eysenck, H. J. & Eysenck, S. B. G. (1975). Manual of the Eysenck Personality Questionnaire. San Diego: Educational and Industrial Testing Service.

Fleeson, W. (2007). Situation-based contingencies underlying trait-content manifestation in behaviour. *Journal of Personality*, 75(4): 825-861.

Fiske, D. (1978). Strategies for personality research The observation versus interpretation of behaviour. San Francisco: Jossey-Bass.

Freud, S. (1943). A General Introduction to Psychoanalysis. Garden City, NY: Garden City Publishing Co.

Goffin, R. D.,&Christiansen, N. D. (2003). Correcting personality tests for faking:Areviewof popular personality tests and initial survey of researchers. *International Journal of Selection and Assessment*, 11, 340–344.

Gigerenzer, G. (1981). Messung und Modellbildung in der Psychologie. München, Reinhardt.

Gough, H.G. (1987). California Psychological Inventory Administrator's Guide. Palo Alto, CA: Consulting Psychologists Press, Inc.

Green & TuU (1978). Research for Marketing Decisions. Englewood Cliffs, NJ: Prentice Hall.

Guilford, J. P. (1954). Psychometric Methods. New York; McGraw Hill.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, 74, 167–184.

Hogan, R., & Hogan, J. (1992). Hogan Personality Inventory manual. Tulsa, OK: Hogan Assessment Systems.

Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decisions: Questions and answers. American Psychologist, 51, 469–477.

Hogan, R. (2005). In defense of personality measurement: New wine for old whiners. *Human Performance*, 18, 331-341.

Holtgraves, T. M. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin*. 30, 161-172.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities (Monograph). *Journal of Applied Psychology*, 75, 581–585.

Hough, L.M. (2003)Emerging trends and needs in personality research and practice: Beyond main effects. In: M.R. Barrick and A.M. Ryan, Editors, *Personality and work: Reconsidering the role of personality in organizations*, Jossey-Bass, San Francisco (2003), pp. 289–325.

Hossiep, R. & Paschen, M. (2003, unter Mitarbeit von O. Mühlhaus). Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP) (2. Aufl.). Göttingen: Hogrefe.

Johnson, C. E., Wood, R. & Blinkhorn, S. E (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology*, 61, 153-162.

Jones, E. E. (1979). The rocky road from acts to dispositions. *American Psychologist*, 34, 107-117.

Judge, T. A., & Kristof-Brown, A. L. (2003). Personality, interactional psychology, and person-organization fit. In B. Schneider, & D. B. Smith (Eds.), *Personality and organizations* (126-161). Mahwah, NJ: Erlbaum.

Jung, C. (1933). Psychological Types. New York: Harcourt.

Kahneman, D., & Tversky, A. (1973) On the psychology of prediction. *Psychological Review*, 80, 237-251.

Lewin, K. (1935). A Dynamic Theory of Personality. New York: McGraw Hill.´

Loewenthal, K. M. (1996). An introduction to psychological tests and scales. London: University College London Press.

Lord, E M. & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley

McDaniel, M. A., Douglas, E. F.,&Snell, A. F. (1997, April). A survey of deception among job seekers. Paper presented at the twelfth annual conference of the Society of Industrial and Organizational Psychology, St. Louis, MO.

Mischel, W. (1968). Personality and assessment New York: Wiley.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, 80, 252-283.

Murray, H. A. (1938). Explorations in Personality. New York: Oxford University Press.

Newcomb, T. (1931). An experiment designed to test the validity of a rating technique *Journal of Education Psychology*, 22, 279-289.

Nunnally, J. C. (1967), Psychometric Theory. New York: McGraw-Hill.

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance*, 11, 245–269.

Ones, D. S.,Viswesvaran, C.,&Reiss, A. D. (1996). Role of social desirability in personality testing for selection: The red herring. *Journal of Applied Psychology*, 81, 660–679.

Orloci, L. (1978) Multivariate Analysis in Vegetation Research (2nd ed.), Junk, The Hague.

Paajanen, G. E., Hansen, T. L., & McLellan, R. A. (1993). PDI Employment Inventory and PDI Customer Service Inventory manual. Minneapolis, MN: Personnel Decisions.

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner &J. S. Wiggins (Eds), *Personality Assessment via Questionnaire*. Berlin: Springer-Verlag.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. Robinson, P. Shaver, & L. Wrightsman (Eds.), *Measures of personality and social psychological attitudes*. San Diego, CA: Academic.

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review*, 111, 781-799.

Reilly, R. R. (1990). An examination of the effects of using behaviour checklists on the construct validity of assessment center dimensions. *Personnel Psychology*, 43, 71-84.

Rogers, C. R. (1947). Some observations on the organisation of personality. *American Psychologist*, 2, 358-368.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment testing and hiring decisions. *Journal of Applied Psychology*, 83, 634–644.

Roth, D. L. & Ingram, R. E. (1985). Factors in the Self-Deception Questionnaire: Associations with depression. *Journal of Personality and Social Psychology*, 48, 243-251.

Rundquist, E. A. (1966). Item and response characteristics in attitude and persotiality measurement. *Psychological Bulletin*, 66, 166-177.

Robertson, I. & Heather, N. (1986). Let's Drink to your Health! Leicester: British Psychological Society.

Sackeim, H. A. (1983). Self-deception, self-esteem and depression: The adaptive value of lying to oneself. In D. Masling (Ed.), *Empirical Studies of Psychoanalytic Theories*. Hillsdale, NJ: Erlhaum.

Saltz, Reece & Ager (1982). Studies of forced-choice methodology: Individuardifferences in social desirability. *Educational and Psychological Measuremen*t, 22, 365-370.

Saville, P., Holdsworth, R., Nyfield, G., Cramp, L. & Mabey, W. (1984). The Occupational Personality Questionnaires. London: SHL.

Shephard, D. (1966). Metric structures in ordinal dazd.. *Journal of Mathematical Psychology*, 3, 287-315.

Shepard, R. N., & Carroll, J. D. (1966). Parametric representation of nonlinear data structures. In P. R. Krishnaiah (Ed.), *Multivariate Analysis* (pp. 561-592). New York, NY: Academic Press.

Shweder, R. A. (1975). How relevant is an individual difference theory of personality? *Journal of Personality*, 43, 455-484.

Shweder, R. A. (1977). Likeness and likelihood in everyday thought: Magical thinking in judgments about personality. *Current Anthropology*, 18, 637-658.

Shweder, R. A., & D'Andrade, R. G. (1979). Accurate reflection or systematic distortion? A reply to Block, Weiss, and Thome. *Journal of Personality and Social Psychology*, 37. 1075-1084.

Simpson, R. H. (1944). The specific meanings of certain terms indicating different degrees of frequency. *Quarterly Journal of Speech*, 1944(30), 328-330.

Spence, K. W. (1956). Behaviour Therapy and Conditioning. New Haven, CT: Yale University Press.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 25-29.

Treisman, A. (1969). Strategies and models in selective attention. *Psychological Review*, 76, 242-299.

Wiggins, J. S. (1986). Epilog. In A. Angleitner & J. S. Wiggins (Eds), *Personality Assessment via Questionnaire*. Berlin: Springer-Verlag.

Yun, G.J., Donahue, L.M., Dudley, N.M. and McFarland, L.A. (2005) Rater Personality, Rating Format, and Social Context: Implications for performance appraisal ratings. *International Journal of Selection and Assessment,* 13, 97–107.

Zavala, A. (1965). The development of the forced-choice rating technique. *Psychological Bulletin*, 63, 117–124.

# 10 Schlussbemerkungen

Die vorliegende Arbeit ist aus einem dreijährigen und stark praxisorientierten Prozess entstanden. Abschliessend sollen die wesentlichen Erkenntnisse in einen grösseren Zusammenhang eingeordnet werden. Die verschiedenen Beiträge haben Erkenntnisse auf mehreren Ebenen geschaffen. Einmal in Form eines Beitrages in der Grundlagenforschung mit der Herausarbeitung neuer Sichtweisen auf die methodische Herangehensweise von Personalbeurteilungsdaten, insbesondere im Hinblick auf Kompetenz-Messverfahren. Dann in der Gegenüberstellung von normativer vs. ipsativer Kompetenzmessung und damit der für die Praxis wichtigen Erkenntnis der bewussten Trennung von Profil- und Profilhöhe bei der Messung von Management Kompetenzen. Und schliesslich auch in einem Beitrag zur Entwicklung eines für die Praxis tauglichen und einfachen Instruments zur Kompetenzmessung.

## 10.1 Methodische Beurteilung der Kompetenzmessung mittels NMDS

Das Schlagwort „Kompetenz“ – insbesondere in Bezug auf einen berufsorientierten Handlungskontext – hat in den vergangenen Jahren die (wirtschafts-)pädagogische sowie die arbeits- und organisationspsychologische Debatte massgebend beeinflusst und mitgeprägt. Profit- und erfolgsorientierte Unternehmungen oder Organisationen sind aufgrund ökonomischer Notwendigkeiten gezwungen, effiziente Personalauswahl und -entwicklungsstrategien zu verfolgen. Für sie ist es mitunter entscheidend, die fähigsten und kompetentesten Mitarbeiter für das Unternehmen aus der Masse der Bewerber oder Arbeitnehmer herauszufiltern, vor allem wenn es um die Besetzung leitender Position geht. Darüber hinaus wäre es für Wirtschaftsunternehmen nahezu grob fahrlässig, wenn die vorhandenen Kompetenzen des „Humankapitals“ nicht ausreichend genutzt und weiterentwickelt würden, da dies eine Konterkarierung hinsichtlich der obligatorischen Unternehmensziele (hier seien insbesondere Effizienz- und Effektivitätsoptimierung benannt) zur Folge hätte. Um den soeben genannten, ehrgeizigen und anspruchsvollen Zielen gerecht zu werden, bedarf es mitunter einer geeigneten und zuverlässigen Messung von Kompetenzen der erfolgversprechenden Kandidaten, um Fehleinschätzungen diesbezüglich zu minimieren. Die Maxime „die richtige Person am richtigen Ort“ sollte hierbei als

Idealvorstellung verfolgt bzw. angestrebt werden, da personale Fehlentscheidungen oftmals weitreichende und schwerwiegende, in der Regel negative Konsequenzen nach sich ziehen, die sich im Zeitablauf sogar potenzieren. Die praktische Relevanz der hier vorgelegten Dissertation, die sich im Besonderen und vor allem auch mit der Thematik der Kompetenzmessung beschäftigt, ist insofern hinreichend dargelegt.

Wissenschaftlich betrachtet, erweist sich die Kompetenzmessung jedoch als schwieriges Unterfangen. Es darf nicht verborgen bleiben, dass durch die Nähe zur Alltagssprache von Kompetenzbegriffen der Nachteil der Diffusität der Bezeichnung (Sarges, 2006) gegeben ist, da begrifflich klar definierte Fachtermini in die jeweilige „Sprachwelt" der Organisation eingebettet und teilweise so modifiziert werden, sodass ein Vergleich zwischen den verschiedenen Kompetenz-Terminologien einzelner Unternehmen schwierig beziehungsweise das Problem der Inkommensurabilität gegenwärtig ist. Insofern scheint es mitunter auch geboten und erforderlich, insbesondere um einen einigermassen homogenen Standard bezüglich der Kompetenzbegriffe und -modelle zu generieren, die Kompetenzkonstrukte mit psychologisch messtechnischem Know-how adäquat und sicher zu erheben. Sarges, der letztlich zu einem Fazit gelangt, dass Kompetenzen respektive das dahingehende Verständnis hinsichtlich eines praxisorientierten Kompetenzbegriffs mehr als alter Wein in neuen Schläuchen ist (Sarges, 2006) und bei weitem über die bewährten, aber teilweise tradierten Konzepte der Anforderungsmerkmale und -analysen hinaus geht, stellt hierbei eine zentrale Forderung auf: Zwar hat sich der Kompetenzbegriff und die Competency-Bewegung im Rahmen der Personalbeurteilung und -entwicklung durchgesetzt, nunmehr sei es jedoch erforderlich, „die Präzision der Messung voran zu treiben und Nachweise der Validität dieser Messergebnisse für die berufliche Leistung zu erbringen" (S. 297).

Die vorliegende Dissertation greift diese Forderung von Sarges auf und bemüht sich um Grundlagenforschung bei der Anwendung verschiedener Kompetenzmess-verfahren. Sozialwissenschaftler wie Erpenbeck und Heyse, welche das Kompetenz-messverfahren KODE[20] entwickelten, haben sich um den Validitätsnachweis ihres Verfahrens bemüht, kamen aber zum Schluss, dass eine Validitätsprüfung nach

[20] KODE steht für Kompetenzdiagnostik und –Entwicklung und gilt als ein wissenschaftlich begleitetes und abgesichertes Verfahren. Es wurde 1996-1998 von Prof. Dr. Heyse und Prof. Dr. Erpenbeck entwickelt.

traditioneller Psychodiagnostik derzeit nicht möglich sei, da KODE kein psychometrisches Messverfahren sei und keine gleichwertigen oder bessere Verfahren existieren würden. Ihr zentrales Argument hierfür ist, dass Kompetenzen sehr kontextbehaftet seien und somit nicht wie überdauernde Persönlichkeitseigenschaften messtheoretisch angegangen werden können. Ich bezweifle die Schwammigkeit der Kompetenzbegriffe sowie die Kontextabhängigkeit der Kompetenzmessung keineswegs – im Gegenteil. Es sind genau diese Eigenschaften der Kompetenzmessung, die neue methodische Ansätzen verlangen, die über die klassische Gütekriterienüberprüfung der Testkonstruktion hinausgehen. Durch die Einführung der Methode der NMDS in der Kompetenzmessung ist eine echte Alternative zu den gängigen in der Testentwicklung üblich angewendeten Verfahren (wie zum Beispiel der Faktorenanalyse) gefunden worden. Das traditionelle Argument lautet, dass die NMDS weitaus geringere Ansprüche an messtheoretische Voraussetzungen als z.B. die metrische Faktorenanalyse stellt: Das nonmetrische multidimensionale Skalierungsmodell war ursprünglich entwickelt worden, gerade weil es so schwierig war, die strengen Annahmen der metrischen Modelle in vorliegenden empirischen Daten zu rechtfertigen (MacCallum, 1974). MacCallum hatte bereits früh erkannt, dass die NMDS im Vergleich zu anderen multivariaten Analysemethoden die schwächsten Annahmen über die Rohdaten treffen muss. So wird einerseits die Skala der Messung als ordinal angenommen und nicht intervallskaliert, und bei den Distanzfunktionen wird nur die Monotoniebedingung und nicht die Linearbedingung gestellt wie etwa bei den metrischen Faktoranalysen oder Korrelationskoeffizienten.

Es gibt jedoch noch einen zweiten, in meinen Augen viel wichtigeren Argumentationsstrang: Neben den schwächeren Annahmen an die Messbedingungen erlaubt die NMDS die Darstellung der Kompetenzen in Relation zu allen anderen (im Kompetenzraum) als auch die Darstellung des Profils jeder einzelnen Führungskraft im Kontext anderer Führungskräfte (im Personenraum). Als Proximitätsmass für die Abbildung der Profile bzw. der Kompetenzen verwendeten wir Pearson-Korrelationen oder Differenzsummen. Unabhängig vom Proximitätsmass liegt mit der NMDS ein statistisches Modell vor, mit dem sich die gesamte Varianz in den Daten abbilden lässt und mit dem man sich mit einem Blick einen Eindruck über die Beziehungen und systematischen Zusammenhänge zwischen einzelnen Personen beziehungsweise

Kompetenzen machen kann. In sonst üblicherweise verwendeten statistischen Methoden haben wir uns zum Beispiel angewöhnt, Unterschiede zwischen Selbst- und Fremdurteile in Spalten- oder Netzdiagrammen oder Korrelationskoeffizienten darzustellen, die oft nur geringe Aussagekraft in Bezug auf das Gesamtbild der Daten besitzen. Hier bietet meiner Ansicht nach die Methodik der robusten multidimensionalen Skalierung den Vorteil, dass sie a) diese Zahlen wieder in eine fassbare Landkarte übersetzt, die intuitiv verständlich ist und mehrere Personen beziehungsweise die unterschiedlichen Sichtweisen mehrerer Beurteiler über verschiedene Personen auf einen Blick zugänglich macht, und dass sie b) aufgrund der Robustheitsfunktion Ausreissern in den Datensätzen keinen Einfluss auf das Gesamtergebnis erlaubt.

Welche potenziellen Anwendungsmöglichskeiten sich daraus für die Praxis ergeben können, wird im letzten Kapitel dieser abschliessenden Diskussion umrissen. Zuerst soll jedoch noch auf die Gegenüberstellung von normativer und ipsativer Kompetenzmessung näher eingegangen werden.

## 10.2 Gegenüberstellung normativer versus ipsativer Messung

Kompetenzinventare und Persönlichkeitsfragebögen werden immer häufiger als Teil von Beurteilungsverfahren in Prozesse der Personalauswahl oder –entwicklung integriert. Im internationalen Vergleich zeigt sich jedoch, dass psychometrische Instrumente, besonders Persönlichkeitsfragebögen, in der Personalauswahl in Deutschland traditionell weit weniger verbreitet sind als in anderen Ländern (z.B. Grossbritannien, Belgien, Niederlande; Ryan, McFarland, Baron & Page, 1999).

Grundsätzlich haben jedoch sowohl die hohe Akzeptanz der Big Five als übergreifendes Persönlichkeitsmodell als auch das breite Angebot an validen Persönlichkeitsfragebögen dazu geführt, dass dieser Prädiktorbereich sich in den letzten Jahren relativ gut etabliert hat (Inceoglu & Bartram, 2007).

Einer der häufigsten Kritikpunkte an die Adresse der Testkonstrukteure und Personalfachleute, die solche psychologischen Testverfahren in der täglichen Personalarbeit einsetzen, ist die Möglichkeit für die Kandidaten der bewussten Verfälschung von Aussagen über die eigene Persönlichkeit bzw. Kompetenzausprägung. Melchers, Klehe, Richter, Kleinmann, König & Lievens (2009) machen aus dieser Not eine Tugend und haben eindrücklich zeigen können, dass die Fähigkeit, sich so darzustellen wie es vom Umfeld verlangt wird (wie zum

Beispiel im Kontext eines Jobinterviews), ein valider Prädiktor für die künftige Leistungserbringung darstellt.

Andere Autoren und Testentwickler haben mit Erfolg das Antwortwortformat der Persönlichkeitsfragebögen verändert, um dadurch die Verfälschungsversuche (Faking) zu reduzieren. (Baron, 1996; Bartram, 2007, Saville & Wilson, 1991; Heggestad, Morrison, Reeve & McCloy, 2006). So sind Forced-Choice Fragebogenformate, welche zu den ipsativen Verfahren zählen, wieder vermehrt in den wissenschaftlichen Diskurs aufgenommen worden. (z.B. Christiansen, Burns & Montgomery, 2005). Diesen Trend haben wir in dieser Arbeit aufgenommen und ein Verfahren entwickelt, dass die Urteilstendenzen bei gängigen Likert-basierten Persönlichkeitsinventaren durch ein kompetenzbasiertes Forced-Choice-Verfahren vermindert. Ipsative Testformate bringen gegenüber Personen, die ein homogenes und damit undifferenziertes Antwortprofil haben, einen gewissen Wahlzwang mit sich. Undifferenziert überhöhte oder erniedrigte normative Testprofile resultieren zum Beispiel bei Akquieszens, Nein-Sage-Tendenz oder bei einer zentralen Tendenz. Im Zusammenhang mit Selbstbeurteilungsverfahren ist vor allem die gut erforschte Selbsterhöhungstendenz ein Problem. Da diese Antworttendenzen mittels des ipsativen Ansatzes reduziert werden können, differenzieren ipsative Testresultate in der Regel besser als ihr normatives Pendant.

Saville & Wilson (1991) schilderten in einer ungewöhnlich anmutenden Art und Weise einen weiteren Vorteil des ipsativen Formats: Leben bestünde an sich aus vielen Entscheidungssituationen, weshalb ipsative Erhebungsmethoden realitätsnäher seien. Der Zwang zur Entscheidung wirke auf diese Weise validitätserhöhend.

Bartram (1996) hingegen warf die Frage auf, ob das in ipsativen Tests verwendete Forced-Choice-Format unterschiedlich bewertet werden müsste. Dies deshalb, weil Probanden in solchen Test-Situationen ihre Wertungen nur rein zufällig vergeben würden. Auch in solchen Konstellationen werden jedoch die wahren Präferenzen, in unserem Fall Stärken und Schwächen, noch eine Einordnung der Alternativen ermöglichen. Es gibt jedoch nach Baron (1996) einen Fall, bei dem das ipsative Format zu Ergebnisverfälschungen führen kann: Wenn die wahren Werte für die Testskalen bei einem Probanden sehr nahe beieinander liegen. Dies vermag ein ipsativer Test nur unzureichend zu erfassen, da der Proband in diesem Fall bei allen Itemgruppen, in denen sich die ähnlich stark präferierten Items aufeinander treffen,

Schwierigkeiten bei der Rangplatzvergabe haben wird. Dies führt dazu, dass der ipsative Test zwischen den problematischen Skalen nur schwer differenzieren kann. Fraglich ist allerdings, ob dieser Fall nicht eher ein Randphänomen darstellt. Baron (1996) sah diesen Randphänomenstatus aufgrund einer von ihr durchgeführten Untersuchung (N=2951) bestätigt. Zudem kann aufgrund der in unserem Test sowohl auf Itemebene (60 Kompetenzfacetten) als auch auf Konstruktebene (15 Kompetenzen) gefundenen hohen Reliabilität (Test-Retest-Reliabilität) davon ausgegangen werden, dass diese Schwierigkeit der Rangordnung nicht sonderlich stark ins Gewicht fällt.

Im weiter oben bereits zitierten Artikel von Christiansen, Burns & Montgomery (2005) wird Kritik an ipsativen Verfahren geübt, nämlich dass durch die relative Natur von Forced-Choice-Items die Axiome der klassischen Testtheorie verletzt würden. Sie schreiben: “Because response choices for each item are generally statements from different traits, choosing one response means that one of another trait is not chosen. When a small number of traits are assessed, being high in elevation on one trait necessitates lower scores on others. This results in mathematical dependency among the trait scales that can create numerous problems for multivariate analyses, such as increased collinearity in regression analyses and improper solutions using factor analytic techniques. Cornwell and Dunlap (1994) present a convincing summary of such criticisms."

Diese Kritik ist unter den drei genannten "If-Bedingungen" absolut gerechtfertigt: 1. Kleine Zahlen von zu vergleichenden Traits (sie erhöhen die statistische Interdependenz), 2. Multivariates Messmodell (setzt Unabhängigkeit der Messungen voraus) und 3. Anwendung auf klassische Testtheorie (hier die Messung der Ausprägung einer Eigenschaft in Relation zur Population). Mit dem von uns vorgeschlagenen Messverfahren und Messmodell gehen wir, wie die einzelnen Kapitel schon gezeigt haben, exakt auf diese drei Aspekte ein:

1. Statt eines Forced-Choice-Verfahrens von wenigen Elementen (z.B. 4 im OPQ32 nach Saville & Holdsworth, 2006) setzen wir als Messverfahren Rangreihen mit 15 Elementen ein. Diese Aufgabenstellung, die eigentlich eher als ein "Forced-Ordering"-Verfahren bezeichnet werden sollte, reduziert die statistische Interdependenz auf ein Minimum: Die Festlegung der Ausprägung eines Kompetenzitems lässt fast uneingeschränkten Spielraum für die anderen Elemente.

(Allein die Zahl der ordinal möglichen Kombinationen steigt von 12 im OPQ32 auf ca. 1'500'000'000'000 Rangreihen aus 15 Objekten.) Die statistische Interdependenz (ein hoher Wert auf der einen Kompetenz hat nicht automatisch einen tiefen Wert in der anderen Kompetenz zur Folge) ist also verringert (wenn auch nicht völlig ausgeschaltet).

2. Auch wenn das Problem der Interdependenz nicht mehr so gravierend ist, wenden wir dennoch keine Multivariaten Analysen, welche die Unabhängigkeit der Messwerte bzw. Faktorstufen erfordern, als Messmodell an. Vielmehr analysieren wir die Daten rein relational mittels NMDS. Die verbleibende statistische Interdependenz ist hier nicht nur kein Problem, sondern als Faktor der Reduzierung von Ausreissern und Inkonsistenzen in den Daten sogar ein Vorteil.

3. Selbstverständlich ist diese Form eines relationalen Messmodells nicht mit der klassischen Testtheorie vereinbar. Diese bestimmt im Fall der Kompetenzmessung die Höhe jeder einzelnen Kompetenz (unabhängig voneinander) in Bezug auf die Population. Das Profil ergibt sich dann a posteriori. Wir argumentieren, dass man lieber die Gesamthöhe eines Profils durch eine zweite, unabhängige Messung vornimmt und dann die Kompetenzen einer Person direkt in Relation zueinander betrachtet und nicht erst in Relation zur Population. Bei dem von uns angewendeten Messverfahren geht es um die intraindividuelle Sicht in Form von relativen Stärken und Schwächen. Dies bedeutet also durchaus eine Abkehr von der klassischen Testtheorie bei der Messung von Kompetenzprofilen. Ohne den entsprechenden Anspruch werden die testtheoretischen Voraussetzungen (wie die statistische Unabhängigkeit, aber auch wie zum Beispiel normalverteilte Daten etc) erst gar nicht benötigt.

Damit haben wir einen Weg gefunden, der alle drei von Christiansen et al. (2005) herausgearbeiteten "If-Bedingungen" ausschaltet. Deswegen können wir mit dieser Kombination aus Messverfahren und Messmodell statistisch einwandfrei arbeiten und so die Kritik von Cornwall und Dunlap (1994) an Forced-Choice-Verfahren für den Bereich der relationalen Kompetenzmessung parieren.

Bezüglich des Vergleichs von normativem und ipsativem Ansatz kann festgehalten werden, dass beide Messkonzepte unterschiedlich auf Störeinflüsse, die sich

zwangsläufig auch bei der Messung von Persönlichkeitseigenschaften bzw. fachübergreifenden Kompetenzen ergeben, reagieren. Die Entscheidung zwischen normativem und ipsativem Ansatz sollte insofern einzelfallbezogen je nach Anwendungskontext getroffen werden. Steht die Vergleichbarkeit im Sinne des absoluten Kompetenzscores oder Leistungsniveau im Vordergrund, ist die normative Methode angebrachter. Sollen Effekte von Verfälschungsversuchen reduziert oder mangelnde Differenzierung bei homogenen normativen Merkmalsprofilen (wie zum Beispiel bei der Personalauswahl zu erwarten) ausgeschlossen werden, ist die ipsative Methode zu empfehlen. Der letztlich anzusetzende Massstab ist stets die erreichbare Validität. Eine Kombination beider Formate wäre eventuell eine Möglichkeit, die Vorteile beider Methoden zu nutzen. Beispielsweise könnte der ipsative Teiltest relative Aussagen zu einem intraindividuellen Stärken- / Schwächenprofil beitragen und der normative Ansatz über das absolute Kompetenz- oder Leistungsniveau Aufschluss geben. In dieser Arbeit konnte gezeigt werden, dass die Information der Profilhöhe (interindividuelle Vergleich gemäss normativem Ansatz) auch durch eine einfache Abfrage des Gesamtlevels erfolgen kann, welches zu den gleichen Ergebnissen führt wie das umständliche Ausfüllen längerer Fragebögen.

Abschliessend möchte ich nochmals betonen, dass das in dieser Arbeit vorgestellte Kompetenzmessverfahren, vor allem im Kontext der Leistungsbeurteilung, einen zentralen Vorteil aufweist, welcher in der Literatur bislang nicht aufgegriffen wurde: Während die gängigen Leistungsbeurteilungsbögen mit mehrstufigen Skalen sowohl die Gesamtprofilhöhe als auch das Profil auf mehreren Verhaltensdimensionen eines Beurteilten abfragen und somit das Arbeitsgedächtnis der Beurteiler extrem stark belasten, reduziert dieses Verfahren die kognitiv hohen Anforderungen an die Beurteiler, was zu exakteren Profilen führt. Basierend auf den in dieser Arbeit gewonnen Erkenntnissen empfehle ich bei Kompetenz-Assessments die Messung der Gesamtprofilhöhe und die Profilinformation strikt zu trennen und in zwei getrennten Schritten durchzuführen. Dass diese Empfehlung durchaus Früchte tragen könnte, betonen die Befunde Bartram, 2005; Guilford & Fruchter, 1973; Mendoza & Mumford, 1987, die zeigen konnten, dass überhöhte Leistungsbeurteilungen mittels gängiger Likert-skalierten Fragebögen zu Varianzreduktion und geringere Differenzierung in Bezug auf verschiedene Urteilsdimensionen (z.B.Kompetenzen) zur Folge haben.

## 10.3 Innovative Anwendungsmöglichkeiten für die Praxis

Das im Rahmen dieser Arbeit entwickelte Selbstbeurteilungsinstrument ist ein innovatives Messverfahren, welches durch zwei Vorteile besticht. Erstens ist es vom ökonomischen Standpunkt her äusserst effizient (die Erhebung eines Profils basierend auf 15 Kompetenzen dauert maximal 2-4 Minuten) und liefert dabei ebenso gute Reliabilitätswerte wie herkömmliche Persönlichkeits- und Kompetenzinventare. Und zweitens eröffnet die Auswertung durch die NMDS neue Möglichkeiten, die Beurteilerdifferenzen ganzheitlich und anschaulich darzustellen.

Unabhängig, ob die interperspektivischen Diskrepanzen wirklich den Dreh- und Angelpunkt in der Vorhersage von Erfolgsmassen darstellen, kann das Ziel, eine höhere Übereinstimmung herzustellen, in der Praxis für sinnvoll erachtet werden. Von den beteiligten Führungskräften und Mitarbeitern werden Ideen erfragt, wie potenzielle Diskrepanzen zwischen Selbst- und Fremdbeurteilern reduziert werden könnten. Da die Diskrepanzen auf unterschiedlichen Wahrnehmungswelten beruhen, die aufgrund Erkenntnisbarrieren zwischen Selbst- und Fremdeinschätzung für den Betroffenen nur schwer nachvollziehbar sind, geht die Überbrückung der Diskrepanzen nur über Kommunikation bzw. konsequentes Einfordern und Geben von Feedback bezüglich der eigenen Wirkung bzw. der Wirkung auf andere. Das Kartenmodell könnte hier eine nützliche Gesprächsgrundlange bieten, um Diskrepanzen ganzheitlich zu visualisieren, darüber zu diskutieren und Wahrnehmungsdifferenzen auszuräumen.

Doch die Anwendungsmöglichkeiten der NMDS gehen weit über die Darstellung von Diskrepanzen in der Selbst- und Fremdbeurteilung hinaus. Interessant wird die Methode vor allem im Hinblick auf ein Matching-Verfahren von Stellen- und Personenprofilen. Dieser Punkt soll abschliessend noch etwas genauer herausgearbeitet werden, da ich der Überzeugung bin, dass in dieser Möglichkeit das grösste Potenzial dieses Messmodells schlummert, welche für die Personalselektion sowie für die Nachfolge- und Karriereplanung in Unternehmen von weitreichender Bedeutung werden könnte. Im Zusammenhang mit der praktischen Anwendung von euklidischen Karten für personale Zwecke ist im vierten Forschungsbeitrag bereits darauf hingewiesen worden, dass ein Matchingverfahren von Job- und Personenprofilen innerhalb einer einzigen Skalierung erfolgen kann. Die Anforderungsprofile werden in Form von Wichtigkeitseinschätzungen der

berufsrelevanten Kompetenzen gebildet, während die Kompetenzprofile durch die Einschätzung des Ausprägungsgrades oder durch die Rangreihenbildung in Analogie zu dem in dieser Arbeit vorgestellten Forced-Choice-Verfahren erstellt werden. Die dadurch gewonnenen Profilinformationen werden sowohl auf Stellenseite als auch auf Personenseite skaliert und basierend auf Korrelationsmatrizen in einen gemeinsamen Raum abgebildet, wobei ähnliche Profile nahe beieinander abgebildet werden und unähnliche weit voneinander entfernt. Dank dieser Methode ergeben sich interessante Ansätze für die gängigen Fragen der Laufbahn- bzw. Nachfolgeplanung. Welche Stelle passt am besten zu einem bestimmten Kandidaten, bzw. welche Kandidaten passen am besten auf eine bestimmte Stelle? Abbildung 7 illustriert schematisch die praktische Anwendung euklidischer Karten in der Personalarbeit.

**Beispiel Laufbahnplanung** **Beispiel Nachfolgeplanung:**

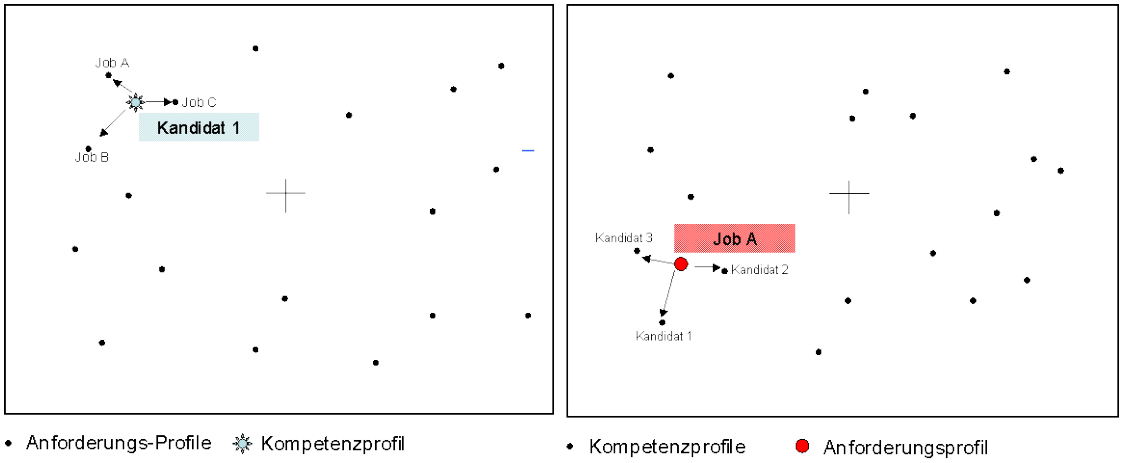



Abbildung Laufbahn- und Nachfolgeplanung mittels Multidmensionaler Skalierung

Voraussetzung für die Skalierung der Daten ist die Verwendung eines gemeinsamen Kompetenzmodells sowohl für die Erstellung von Anforderungsprofilen als auch für die Messung von Kompetenzprofilen. Bei integrierten Kompetenzmanagement-Ansätzen wie sie etwa Hilb (2008) vorschlägt, ist dies auch der Fall. Zudem bleibt gemäss den in meiner unveröffentlichten Lizentiatsarbeit gewonnenen Erkenntnissen (Zuber, 2005) darauf hinzuweisen, dass bei den Wichtigkeitseinschätzungen der Anforderungsprofile als auch bei den Kompetenzprofilen, welche jeweils auf Likert-Skalen erhoben werden, eine Punktebegrenzung vorgegeben werden muss, damit die Vergleichbarkeit der beiden Profilinformationen gegeben ist. Bei einem Forced-Choice Verfahren im Sinne einer Rangreihenbildung besteht diese Gefahr nicht.

Da jedoch in diesem hier vorgestellten Ansatz die Profilhöhe, d.h. das absolute Level, sowohl auf Stellenseite als auch auf Personenseite nicht berücksichtigt wird, muss diese Information auf anderen Wegen erhoben werden. Auf Stellenseite könnte dabei ein Funktionsbewertungssystem Abhilfe schaffen, z.B. nach der Hay-Methode[21], wo verschiedene Stellen gemäss ihrem spezifischen Anforderungsniveau eingestuft werden. Solche Funktionseinstufungen sind in den meisten Unternehmungen bereits vorhanden, da sie die Voraussetzung für anforderungsgerechte Entgeltungssysteme sind. Auf Personenseite kann z.B. das Ausbildungsniveau, die Anzahl Jahre an Berufserfahrung oder das bisherige Gehalt (in Zusammenhang mit dem Alter der Person) Aufschluss über die Profilhöhe geben. Bei der Stellenbesetzung geht man ohnehin davon aus, dass die potenziellen Kandidaten auf einem ähnlichen Kompetenzniveau sind und die Mindestanforderungen der Stelle erfüllen. Mittels der NMDS kann jedoch nun noch genauer spezifiziert werden, welcher Kandidat basierend auf dem Kompetenzprofil auf die Stelle am besten passt. Es ist jedoch nochmals darauf hingewiesen, dass man beim Vergleich von Kompetenz- und Stellenprofilen mittels NDMS die jeweilige Profilhöhe im Vorfeld bestimmt hat und somit nicht Äpfel mit Birnen vergleicht.

Der Autor wünscht sich noch weitere Forschung auf diesem Gebiet, indem die Passung zwischen Anforderungs- und Kompetenzprofil mittels NDMS einem externen Kriterium unterzogen wird, um zu zeigen, dass die Vorhersage mittels dem NMDS matching zu den gewünschten Ergebnissen führt. Dieses externe Kriterium könnte die prognostische Validität betreffen, in dem man sowohl auf Personen- als auch auf Stellenseite untersucht, ob die Person nach mehreren Jahren auf dem gemäss NMDS-Lösung am besten passenden Job erfolgreich war, zum Beispiel mittels Korrelationsmessung mit weiteren Beförderungen oder Gehalt. Bei einer genügend grossen Anzahl solcher validen Prognosen könnte die NMDS in der Management-Diagnostik einen hohen Stellenwert erreichen.

[21] Die Hay-Methode ist nach der HeyGroup benannt. Die HayGroup ist eine weltweit tätige Unternehmensfirma, die sich auf den Bereich Human Resource Management spezialisiert hat und ihre Kernkompetenzen vor allem im Bereich der Stellenbewertung und dem Gehaltsmanagement aufweist (vgl. HayGroup o.J.a. sowie www.haygroup.com).

Ich möchte meine Arbeit abschliessen, indem ich ein mögliches praxisorientiertes Forschungsprojekt vorstelle, das die oben skizzierten Gedanken aufnimmt, jedoch die Anwendungsmöglichkeiten auf mehrere Unternehmen und Personengruppen erweitert.

Die zentralen Ideen eines solchen Forschungsprojekts, das die Konstruktion eines innovativen Personalbeurteilungsinstruments beinhaltet, soll an dieser Stelle nur skizziert werden. Im schlechtesten Fall dienen diese Ideen lediglich einer gedanklichen Inspiration - im besten Fall vermögen sie ein konkretes Forschungsprojekt auszulösen.

Die Grundlage des im Rahmen dieser Diskussion grob umrissenen Personalbeurteilungs-instruments basiert auf der Konstruktion einer für die Persönlichkeits- bzw. Kompetenztypologie repräsentativen NMDS-Karte und dessen Aussagekraft für daraus ableitbare Personalmassnahmen. Ziel des Vorschlags ist es, eine NMDS-Karte zu bauen, in dem alle möglichen Typen von Persönlichkeits- oder Kompetenzprofilen in einem zweidimensionalen Raum gemäss ihrer relationalen Ähnlichkeitsstruktur abgebildet sind. In diese „Standard-Karte“ könnte man dann sowohl reale Personen- als auch reale Stellenprofile in die Korrelationsmatrix einzeln hineinrechnen und in der Standard-Karte abbilden (z.B. mittels externem Unfolding, eine Einführung dazu gibt das Buch von Borg & Groenen, 2005). Mit der nötigen Softwareunterstützung, welche die NMDS-Karte abbildet und die darin enthaltenen Punkte per Mausklick als Profile hervorheben lässt (inkl. Pop up von CV und Stellenbeschreibung), liessen sich Fragen der Passung von Stelle und Person auf einen Blick ablesen.

Zur Konstruktion einer dazu notwendigen „Standard-Karte“ müsste man eine genügend grosse Anzahl an Personen (einige hundert) gemäss einer nach Branchenzugehörigkeit, Fachbereich, Alter und Managementstufe geschichteten Stichprobe auswählen und mittels ipsativer Messung, z.B. mit dem in dieser Arbeit vorgestellten Forced-Choice-Verfahren, von jeder Person Kompetenzprofile erheben.

Nun interessiert die Frage, wie diese grosse Anzahl an Kompetenzprofilen zusammenhängen und strukturiert sind. Wir gehen davon aus, dass die Variation in solchen Profilen nicht unendlich gross ist, sondern dass sich bestimmte Gruppen von Profiltypen herausbilden. Ein geeignetes Verfahren, welches Gruppen von ähnlichen Profilen zusammenschliesst, ist eine Clusteranalyse, welche die Grobstruktur der Varianz in den Korrelationen kategorial abbildet. Der Fokus bei der Interpretation der

Clusteranalyse müsste auf der Anzahl Cluster, der Anzahl Profile pro Cluster sowie der jeweiligen Entfernung zwischen den verschiedenen Clustern liegen. Nehmen wir einmal an, wir hätten ca. 30 Cluster mit jeweils 10 Profilen, wobei einzelne Cluster natürlich mehr und andere weniger Profile enthalten könnten. Ziel wäre es nun, pro Cluster prototypische, reale Profile zu identifizieren, die den Cluster am besten repräsentieren. Ein Protoypen-Mass könnte zum Beispiel die höchste mittlere Korrelation zu allen anderen Profilen dieses Clusters sein. In einem nächsten Schritt würde man je nach Grösse des Clusters 1- 3 prototypische Profile in eine neue Korrelationsmatrix bringen und mittels NMDS skalieren. Wichtig dabei ist, dass man die Profile nicht mittelt, sondern die realen Profile gemäss der soeben geschilderten Prototypenidentifikation nehmen würde, um sie in der Standard-Karte abzubilden. Das Resultat wäre eine Karte von 60 bis 70 Kompetenzprofilen, die gemäss ihrer Relation zu allen anderen Profilen an einen bestimmten Ort in der Karte zu liegen kommen und inhaltlich für einen ganz spezifischen Kompetenztyp stehen. Wie man diese einzelnen Kompetenztypen benennen würde, um somit eine schnelle Matching-Interpretation zu ermöglichen, ist keine leichte Aufgabe und bedarf guter Kenntnisse der Persönlichkeitspsychologie. Doch diese Aufgabe scheint uns nicht sonderlich schwieriger als die Interpretation von Faktorladungen bei der Bestimmung und Benennung der Hauptkomponenten bei der Fragebogenkonstruktion traditioneller Persönlichkeitstests.

## 10.4 Literatur

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, 69, 49–56.

Bartram, D. (1996). The relationship between ipsatized and normative measures of personality. *Journal of Occupational & Organizational Psychology*, 69, 25–39.

Bartram, D. (2005). The Great 8 competencies: A criterioncentric approach to validation. *Journal of Applied Psychology*, 90, 1185-1203.

Bartram, D- (2007). Increasing Validity with Forced-Choice Criterion Measurement Formats. *International Journal of Selection and Assessment*, 15, 263-277.

Borg, I. & Groenen, P. (2005). Modern Multidimensional Scaling: Theory and Applications (Second Edition), Springer, New York, NY, 2005.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance*, 18, 267–307.

Dunlap, W. P., & Cornwell, J. M. (1994). Factor analysis of ipsative measures. *Multivariate Behavioral Research*. 29, 115-126.

Guilford, J. P. & Fruchter, B. (1973). Fundamental statistics in psychology and education (5th ed.). New York: McGraw-Hill.

HayGroup (Hrsg.) (o.J.a). The Hay Guide Chart – Profile Method of Position Evaluating, o.O.o.J.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology*, 91, 9–24.

Hilb, M. 2008. Integriertes Personalmanagement. Ziele -Strategien Instrumente 18. Auflage, Heidelberg: Luchterhand.

Inceoglu, I. & Bartram, D. (2007). Die Validität von Persönlichkeitsfragebögen: Zur Bedeutung des verwendeten Kriteriums. *Zeitschrift für Personalpsychologie*, 6, 160-173.

MacCallumn, R. (1974). Relations between Factor Analysis and Multidimensional Scaling. *Psychological Buelletin*, 81, 8, 505-516.

Melchers, K. G., Klehe, U.-C., Richter, G. M., Kleinmann, M., König, C. J., & Lievens, F. (2009). "I know what you want to know": The impact of interviewees' ability to identify criteria on interview performance and construct-related validity. *Human Performance*, 22, 355-374.

Mendoza, J. L. & Mumford, M. (1987). Correction for attenuation and range restriction on the predictor. *Journal of Educational Statistics*, 12, 282-293.

Ryan AM, McFarland L, Baron H, Page R. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. *Personnel Psychology*, 52,359-391.

Sarges, W. (2006). Competencies statt Anforderungen – nur alter Wein in neuen Schläuchen? In H.-C. Riekhof (Hrsg.), *Strategien der Personalentwicklung* (6. Aufl.; S. 133-148). Wiesbaden: Gabler.

Saville, P., & Willson, E. (1991). The reliability and validity of normative and ipsative approaches in the measurement of personality. *Journal of Occupational Psychology*, 64(3), 219-238.

Saville and Holdsworth Ltd. (1993). Best Practice in the Management of Psychometric Tests: Guidelines for Developing Policy. Surrey: Saville and Holdsworth Ltd.

SHL. (2006). OPQ32: Technical Manual. Thames Ditton, UK: SHL Group plc.

Zuber, T., Matthys, A., & Läge, D., (2005). Anforderungsorientierte Managementtypologie. AKZ-Forschungsbericht Nr. 18. Zürich: Angewandte Kognitionspsychologie